WESTERN SYDNEY
UNIVERSITY

Characterising heterogeneity in the T-cell Acute Lymphoblastic Leukaemia Jurkat cell line in the context of the *TAL1* locus

Presented by:

Raisa Hasan

BMSc – Biomedical Major

A thesis submitted in fulfillment of the requirement of

Master of Research

Principal Supervisor: Dr. Graham Jones

Western Sydney University, Australia

November 2019

Acknowledgements

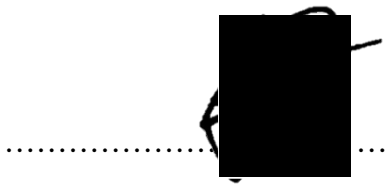
I want to extend my gratitude to my supervisor, Dr. Graham Jones, for the never-ending support and guidance over the last two years of my degree. Thanks to you, I have found a passion for science and teaching which I didn't know I had. Thank you for providing me with advice and skills that extend beyond the lab and will benefit me for the rest of my life.

I would also like to thank my parents who have taught me the meaning of hard work, and for providing me with the opportunities to achieve any goal I set for myself. Thank you for believing in me when I doubted myself the most. I would also like to thank my sisters, Ermina and Ameera. You two have been my number one supporters since the beginning. You continue to teach me and help me grow despite being so much younger than me. I will continually strive to be the best role model possible for you both.

Thank you to my lab buddies from day one, Emily, Krisel and Elyssa. We all started together on this journey and we will soon be on our separate paths, but these two years wouldn't have been enjoyable without you guys. I would also like to thank all of my colleagues in the post-grad room, your advice and encouragement have helped me grow and thrive over these last two years. Thank you to the laboratory technical staff for always making sure everything was working and also providing endless support. I want to thank my friends outside of the university for being always being proud of me and cheering me on even when I am too busy to see you. Lastly, thank you Brandon, you have influenced me to grow in so many ways and I wouldn't be the person I am today without you.

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.



Raisa Hasan

November 2019

Table of Contents

Acknowledgements ..... *i*

Table of Contents ..... *iii*

List of Tables ..... *vii*

List of Figures ..... *xi*

Common Abbreviations and Gene Names ..... *xv*

Abstract ..... *xvi*

Chapter 1 - Background *1*

1.1. Cell Differentiation Programs can be Hijacked by Cancer **1**

1.2. Acute Lymphoblastic Leukaemia (ALL) **1**

1.3. A Subtype of ALL: T-Cell Acute Lymphoblastic Leukaemia (T-ALL) **2**

1.3.1 The *TAL1* gene in T-ALL **6**

1.4. Epigenetic Landscapes: DNA Methylation **10**

1.5. Epigenetic Landscapes: Regulatory Elements - Enhancers **13**

1.5.1 Enhancers within T-ALL – Jurkat Super Enhancer 19

..... **21**

1.7. Heterogeneity within Cancer Cell lines **22**

..... **25**

1.8 Hypothesis **26**

1.9. Objectives and Aims **26**

Chapter 2 - Characterisation of Jurkat Clonal Cell Lines *27*

2.1. Introduction **27**

2.2. Methods and Materials **29**

2.2.1. Jurkat Cell Thawing and Culturing 29

2.2.2. Jurkat Cell Clonal Population Generation 29

2.2.3. Jurkat Clonal Populations 30

2.2.4. Carboxyfluorescein Succinimidyl Ester (CFSE) assay 30

2.2.5. Propidium Iodide (PI) Staining 32

2.2.6. CFSE Analysis 32

2.2.7. RNA isolation 35

2.2.8. cDNA Preparation 35

2.2.9. Primer Design 36

2.2.10. Quantitative Analysis of Gene Expression (qPCR) 37

2.2.11. qPCR Gene Expression Analysis 37

2.3. Results **38**

2.3.1. Optimisation of CFSE assay using the Jurkat Clonal Populations 38

2.3.2. Analysis of proliferation index between Jurkat parental cell lines and the clonal cell lines 45
..... 46
2.3.3. Analysis of the proliferation index between different Jurkat clonal cell lines 47
2.3.4. Subgrouping of Jurkat Clones based on a parental Jurkat cell line (P0) outlier analysis 47
2.3.5. Optimisation of qPCR Gene Expression assay 53
2.3.6. Clustering of Jurkat clonal cell lines based on gene expression of the TAL1 CRC genes 54
2.3.7. Permutation Analysis for *TAL1* CRC genes 57
2.3.8. Differences between Jurkat clonal cell line gene expression between passages 59
2.3.9. Statistical testing of Jurkat gene expression groups 61
2.3.10. Gene Expression and its relationship with the Jurkat cell line proliferation phenotypes 62

2.4. Conclusion 63

Chapter 3 - Bioinformatic analysis of the TAL1 locus 65

3.1. Introduction 65

3.2. Methods 67
3.2.1. Source and type of ChIP-seq Data 67
3.2.2. Quality Control Processing of ChIP-Seq Files 67
3.2.3. FASTQ sequencing file genome alignment, conversion and peak calling 68
3.2.4. ChIP-Seq Differential Binding Analysis (DBA) 70
3.2.5. Processed data visualisation in the UCSC genome browser 71

3.3. Results 74
3.3.1. Differential Binding Analysis of ChIP-seq data between multiple cell lines 74
3.3.1.a DBA analysis of DNAse1 Hypersensitivity within the TAL1 locus 80
3.3.1.b DBA analysis of H3K27ac within the TAL1 locus 83
3.3.1.c DBA analysis of H3K4me3 within the TAL1 locus 86

3.4. Conclusion 89

Chapter 4 - DNA methylation and genetic variation across the TAL1 locus 90

4.1. Introduction 90
4.1.1. DNA methylation 90
4.1.1.a. Methylation-Sensitive Restriction Endonuclease (MSRE) assay 91

4.1.2. Nanopore sequencing 92

4.2. Methods 93
4.2.1. Bioinformatic analysis of the DNA methylation landscape across *TAL1* 93
4.2.2. gDNA isolation of Jurkat Clonal Populations 95
4.2.3. MSRE Protocol 97
4.2.4. MSRE assay analysis 98
4.2.5. Amplicon generation 99
4.2.6. Nanopore library preparation and sequencing 102
4.2.7. Nanopore Sequencing Analysis 104

4.3. Results 107
4.3.1. MSRE Optimisation 107
4.3.2. The DNA methylation landscape of *TAL1* 112
4.3.3. MSRE assay on Jurkat parental and clonal cell lines (P0, C11 and C4) for regions across the *TAL1* locus 116
4.3.4. Amplicon Generation Optimisation for Nanopore Sequencing 123
4.3.5. Nanopore sequencing of the Jurkat parental and clonal cell lines across *TAL1* . 126
4.4. Conclusion 134
Chapter 5 - Discussion and Future Directions 137
5.1. Hypothesis 137
5.2. Summary of Results 137
5.2.1. Phenotypic and Transcriptional Profiles of Jurkat parental and clonal cell lines 137
5.2.2. The understanding of the intergenic *MuTE* insertion enhancer within Jurkat cell lines 138
5.2.3. The intragenic landscape of *TAL1* within Jurkat parental and clonal cell lines ... 139
5.3. Future Directions 142
5.5. Conclusion 146
Chapter 6 - References 147
Chapter 7 - Appendix 159
7.1 – Chapter 2: Supplementary Figures, Tables and Command Lines 159
7.1.1 Supplementary Figures 159
..... 159
7.1.2 Supplementary Tables 159
7.1.2.a qPCR primer sets and amplicon sequences (Methods: 2.2.9) 162
7.1.2.b Pheatmap Command Line – R programming language 167
7.1.2.c GSALightning Command Line – R Programming language 168
7.2 – Chapter 3: Supplementary Figures, Tables and Command Lines 170
7.2.1 Supplementary Figures 170
..... 171
..... 171
..... 172
7.2.2 Supplementary Tables 174
7.2.3 Chapter 3 - Command Lines Used 179
7.2.3.a FastP Code - Python 179
7.2.3.b Bowtie2 Code – Python 180
7.2.3.c Samtools Code for .bam conversion, sorting and indexing - Python 180
7.2.3.d DiffBind code – R programming language 181
7.3 – Chapter 4: Supplementary Figures, Tables and Code 184
7.3.1 Supplementary Figures 184
..... 187
7.3.2 Supplementary Tables 189
7.3.2.a MSRE primer set and amplicon sequences (Methods: 4.2.1) 189

7.3.2.b Sequencing amplicon sequences (Methods: 4.2.5) 192
7.3.3 Chapter 4 - Command Lines Used 205
7.3.3.a qcat code - Python 205
7.4.3.b Minimap2 – Python 205
7.4.3.c Bcftools – Python 205

List of Tables

| Table 1.1. Summary of gene categories and targets that have genetic rearrangements or mutations that are commonly found within T-cell acute lymphoblastic leukaemia (T-ALL) [1]. 3

| Table 2.1. Sequences for Forward and Reverse Primers of the genes, *TAL1*, *GATA3*, *MYB*, *RUNX1* and *NFYB*. 36

| Table 2.2. Mann-Whitney U test statistical test results of proliferation index (PI) of parental Jurkat cell lines (P0 and P00) and Jurkat clones (C1-C6, C8-C11). 48

| Table 2.3. List of Jurkat cell lines categorised by their proliferative ability from the boxplot outlier analysis (Fig 2.7) (see also Methods: 2.2.6 for calculation method). 50

| Table 2.4. Post-Hoc test using Wilcoxon-Rank Sum Test results of pairwise-comparisons between passages 1, 5 and 9 (P1, P5, P9, respectively). Significance determined by $p < 0.05$ (*). 53

| Table 2.5. List of Jurkat cell lines and their passage number for the gene expression groups identified through Euclidean Distance and Hierarchical Clustering of fold expression patterns. 60

| Table 2.6. Permutation q-values amongst gene sets at passage 1, 5, 9 and all passages combined determined by GSALightning R package. 61

| Table 2.7. Kruskal-Wallis test statistics of differences between gene expression groups ($TAL1^+$, $GATA3^+$ and $RUNX1/MYB^+$) for expression of the genes *TAL1*, *GATA3*, *RUNX1* and *MYB*. 62

| Table 3.1. List of cell lines tested for consensus peaks in replicate samples using 'DiffBind' consensus peak overlap. 73

| Table 3.2. DBA Analysis of the *TAL1* locus. 81

| Table 4.1. List of MSRE primers designed to target CpG dinucleotide sites within the *TAL1* locus. 94

| Table 4.2. Averaged duplicates for *TAL1* gene expression (Chapter 2) for Jurkat clonal and parental cell lines at passage 1 and 9 and arranged based on relative *TAL1* expression[2]. 96

| Table 4.3. Parameters used for gDNA digestion with endonucleases HpaII, MspI and mock digested gDNA. 97
|Table 4.4. Primers designed for MinION Nanopore sequencing for targeted regions across the *TAL1* locus. 100
| Table 4.5. PCR cycling conditions for *TAL1* amplicons using the Q5 High-Fidelity polymerase. 101
| Table 4.6. Digestion optimisation conditions with endonucleases HpaII and MspI for Jurkat cell line gDNA. 107
| Table 4.7. Previous 2-step qPCR cycling conditions changed to the 3-step qPCR cycling conditions for a more sensitive MSRE assay. 109
| Table 4.8. DNA methylation statuses of Jurkat parental and clonal cell lines (P0, C11 and C4) as found by the MSRE assay for MSRE regions 2-5 and *TAL1* expression at passages 1 and 9. 121
| Table 4.9. Called genotypes for finalised Jurkat cell line SNVs and Indels with a list of co-localising GWAS/common SNPs from 1000 Genomes Project and gnomAD database and their alternative allele and allele frequencies. 130
| Table 7.1. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Holm FWER method between all Jurkat cell lines at passage 1 for averaged proliferation index 159
| Table 7.2. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Benjamini-Hochberg FDR method between all Jurkat cell lines at passage 1 averaged proliferation index 160
| Table 7.3. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Holm FWER method between all Jurkat cell lines at passage 5 averaged proliferation index..... 160
| Table 7.4. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Benjamini-Hochberg FDR method between all Jurkat cell lines at passage 5 averaged proliferation index 161
| Table 7.5. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Holm FWER method between all Jurkat cell lines at passage 9 averaged proliferation index..... 161
| Table 7.6. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Benjamini-Hochberg FDR method between all Jurkat cell lines at passage 9 averaged proliferation index 161

| Table 7.7. List of primers sets for qPCR analysis of genes *TAL1*, *GATA3*, *MYB* and *RUNX1* 162
| Table 7.8. Friedman Test results for averaged gene expression of TAL1 CRC genes at passages 1, 5 and 9 ($p<0.05$). 165
| Table 7.9. Wilcoxon Rank Sum Two-Tailed Test as Post-Hoc for Friedman test (Table 7.8) between passages 1 and 5, 1 and 9 and 1 and 5 for averaged TAL1 CRC gene expression 165
| Table 7.10. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Holm FWER method for *TAL1* gene expression between gene expression groups 166
| Table 7.11. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Benjamini-Hochberg FDR method for *TAL1* gene expression between gene expression groups 166
| Table 7.12. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Holm FWER method for *GATA3* gene expression between gene expression groups 166
| Table 7.13. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Benjamini-Hochberg FDR method for *GATA3* gene expression between gene expression groups 166
| Table 7.14. Spearman's Rho Correlation test p-values for the $GATA3^{+}$ group and gene expression of TAL1 CRC genes..... 166
| Table 7.15. Spearman's Rho Correlation test p-values for the $TAL1^{+}$ group and gene expression of TAL1 CRC genes..... 166
| Table 7.16. Spearman's Rho Correlation test p-values for the $RUNX1/MYB^{+}$ group and gene expression of TAL1 CRC genes..... 167
| Table 7.17. Spearman's Rho Correlation test p-values for all Jurkat parental and clonal cell lines and gene expression of TAL1 CRC genes 167
| Table 7.18. Spearman's Rho Correlation test p-values for proliferation index (PI) correlation to TAL1 CRC gene expression over all passages, passage 1, 5 and 9 for all Jurkat parental and clonal cell lines 167
| Table 7.19. Spearman's Rho Correlation test p-values for proliferation index (PI) correlation to TAL1 CRC gene expression for all passages for gene expression groups 167

| Table 7.20. List of DNAse1 Hypersensitivity files downloaded from the ENCODE project used for differential binding analysis 174
| Table 7.21. List of H3K27ac ChIP-seq data from the ENCODE Project and other studies (GEO Accession and SRA accession) 175
| Table 7.22. List of H3K4me3 ChIP-seq data from the ENCODE Project and other studies (GEO Accession and SRA Accession) 176
| Table 7.23. List of Input Files for ENCODE and other studies ChIP-seq data (GEO Accession and SRA Accession) 177
| Table 7.24. Phred Scores of ChIP-seq data for the cell line, Jurkat and primary cell lines, Th0, Th1 and Th2 for H3K27ac, H3K4me3 and Inputs from FastQC and after FastP trimming. 177
| Table 7.25. Example of the top 50 results from DiffBind Differential Binding Analysis Granges Object for primary T-cells (Th0 and Th1) and the Jurkat cell line for DNAse1 Hypersensitivity (Cell lines from Table 7.20-23). 178
| Table 7.26. List of MSRE Primers, Primer3 outputs and sequences of MSRE primer amplicons from Chapter 4 (Methods: 4.2.1) 189
| Table 7.27. Kruskal-Wallis results for DNA methylation of T-ALL, B-ALL and myeloid leukaemia cell lines for CpG Islands 1 and 2 within the *TAL1* locus. 199
| Table 7.28. PCR cycling conditions as conducted for Q5 high-fidelity polymerase (per manufacturer's instructions). 200
| Table 7.29. Table of Proxy SNPs of tag SNPs that were found to co-localise Jurkat SNVs with linkage disequilibrium coefficient (r^2>0.8), coordinates, alleles, distance from tag SNP, r^2 coefficient, correlated alleles and RegulomeDB values. 200
| Table 7.30. SNV3 Predicted JASPAR 2020 Transcription Factor Binding, Score and bp from start of the tested sequence for both the Jurkat alternate allele and the reference allele. 202
| Table 7.31. SNV5 Predicted JASPAR 2020 Transcription Factor Binding, Score and bp from start of the tested sequence for both the Jurkat alternate allele and the reference allele. 203

List of Figures

| Figure 1.1. T-Cell Acute lymphoblastic leukaemia (T-ALL) development occurs within the lymphoid differentiation process for T-cells. 4
| Figure 1.2. TAL1 core regulatory circuit (CRC) positive auto-regulatory loop. 8
| Figure 1.3. Model of TAL1-induced transformation in late cortical T-ALL 10
| Figure 1.4. DNA methylation patterns amongst haematopoietic stem cells (HSCs) and the changes of DNA methylation during differentiation within somatic cells. 12
| Figure 1.5. Open chromatin is identified by H3K27ac and H3K4me3 and DNAse1 hypersensitivity (DHS). 16
Figure 1.6 (A-B). Model for transcription at intergenic and intragenic enhancers and the influence of DNA methylation at intragenic and intergenic enhancer sites for transcription 19
| Figure 1.7. Differences between typical enhancers and super enhancers. 20
| Figure 1.8. The organisation of *TAL1* loci within the Jurkat T-ALL cell line displaying an acquired super-enhancer upstream of the *TAL1* gene (chr1:47,705,000). 21
| Figure 1.9. Model of testing cancer cell line heterogeneity through the generation of clonal populations from single cells. 25
| Figure 2.1. Identification of Dead Cells using Propidium Iodide 40
| Figure 2.2. Flow cytometry histogram plots of CFSE fluorescence at 0 and 72 hours post-labelling in the presence of 5% FBS. 41
| Figure 2.3. Flow-cytometry dot plots demonstrating the forward scatter (FSC) (x-axis) and side scatter (SSC) (y-axis) of Jurkat Cells to distinguish cell populations. 42
| Figure 2.4. Flow cytometry histogram plots of CFSE stained Jurkat Clone 2 at cell concentrations of $2x10^5$, $1x10^5$ and $5x10^4$/well. 43
| Figure 2.5. Proliferation indices for Jurkat parental and clonal cell lines. 46
| Figure 2.6. Boxplot display of proliferative index (PI) duplicate values per clone (n=6) at passages 1, 5 and 9. 48
| Figure 2.7. Outlier analysis of proliferative index for parental and clonal cell lines. 49

| Figure 2.8. Boxplot display of proliferative index (PI) data distribution of all clones at passage 1, 5 and 9. 52
| Figure 2.9. (A-D) Primer Efficiencies for TAL1 CRC genes. 55
| Figure 2.10. (A-D) Melt Curve analysis of TAL1 CRC primers. 56
| Figure 2.11. Euclidean Distance Hierarchical clustering of TAL1 CRC gene expression amongst Jurkat cell lines at passages 1, 5 and 9. 58
| Figure 3.1. Venn diagram of H3K27ac enriched peaks within Th2 cell line replicates (red and green circles). 75
| Figure 3.2. Correlation heatmap based on read counts performed by the R package, DiffBind for H3K27ac ChIP-seq data. 76
| Figure 3.3. Differential binding analysis (DBA) heatmap based on contrasts of H3K27ac enrichment between lymphocyte cell lines (Jurkat, CD20RO01794, Th1 and Th2) and non-immune cell lines (HMEC and NHEK). 78
| Figure 3.4. UCSC Genome browser display of the regulatory element markers, DHS, H3K27ac and H3K4me3 across the *TAL1* locus and with localising regions of differential binding. 79
| Figure 4.1 Digested gDNA with endonucleases HpaII, MspI and mock digested at digestion concentrations of 150ng and 300ng and qPCR gDNA concentration with 10ng or 20ng. 108
| Figure 4.2. (A-B) Melt Curve plot of duplicate amplicons generated through MSRE optimisation. 110
| Figure 4.3. (A-E) MSRE Primer Efficiencies. 111
| Figure 4.4. Identification of MSRE Target Regions. 114
| Figure 4.5. Display of CCLE tested CpG Island sites within the *TAL1* gene and percent methylation. 115
| Figure 4.6. MSRE Analysis of the *TAL1* locus in Jurkat Cell Lines. 118
| Figure 4.7. (A-B). Optimisation of PCR amplicons (Seq_1 - Seq_6) with 75ng of gDNA using the touch-down PCR cycling conditions. 124
| Figure 4.8. Example of the metrics provided by European Galaxy database 'Nanoplot' for nanopore sequencing data. 125

| Figure 4.9. IGV detection of a SNV and 12-bp insertion after Nanopore Sequencing of *TAL1*. 127
| Figure 4.10. Mapping of Jurkat cell line genetic variants within the *TAL1* locus identified using nanopore sequencing and co-localising published SNPS. 129
| Figure 4.11. Linkage disequilibrium patterns of co-localising published SNPs with Jurkat SNVs identified through Nanopore Sequencing within the *TAL1* locus. 132
| Figure 4.12. Prediction of Jurkat SNV functional relevance within regulatory elements. 133
| Figure 7.1. Flow cytometry histogram plots of CFSE stained Jurkat Clone 2 at cell concentration of $2x10^4$/well 48-hour growth in 10% FBS 1% PSF in RPMI 1640 with 0.5µM of CFSE. 159
Figure 7.2. Example of the quality score across all bases for Illumina 1.9 sequencing of Jurkat H3K27ac replicate 1 (GEO Accession: GSM1431908) using the FastQC program (Method:3.2.2). 170
| Figure 7.3. Example of MACS2 outputs of Illumina 1.9 ChIP-seq data peak model analysis and cross-correlation analysis from the Jurkat H3K27ac replicate 1 data (GEO Accession: GSM1431908). 171
| Figure 7.4. H3K27ac read depth across the TAL1 locus for all replicates for cell lines, Jurkat, DND41, HMEC and NHEK, and primary cell lines Th1, Th2, CD20+ B-cells and MonoCD14+. 172
| Figure 7.5. DNAse1 hypersensitivity (DHS) read depth displayed in the UCSC genome browser for the all replicates of Jurkat, CD20R017794, MonoCD14, $CD4^+$ Naïve Wb11970640, SAEC, 8988T and Th1 cell lines across the *TAL1* locus. 173
| Figure 7.6. An example of GC content distribution of the sequence amplified by Seq_1 primer set within 30bp windows. 184
| Figure 7.7. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV10 – chr1:47,705,674 184
| Figure 7.8. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV8 – chr1:47,704,240 185
| Figure 7.9. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV9 – chr1:47,704,674 185

| Figure 7.10. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV7 – chr1: 47,703,613 186
| Figure 7.11. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV6 – chr1: 47,697,125 186
|Figure 7.12. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV5 – chr1:47,695,997 187
| Figure 7.13. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV4 – chr1: 47,693,220 187
| Figure 7.14. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV3 - chr1: 47,693,116 188
| Figure 7.15. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV2 – chr1: 47,692,786 188
| Figure 7.16. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV1 – chr1:47,692,281 189

Common Abbreviations and Gene Names

Abbreviation	Definition
CD4+	Cluster of Differentiation 4
CFSE	Carboxyfluorescein Succinimidyl Ester
CRC	Core Regulatory Circuit
DHS	DNAse1 Hypersensitivity
GATA3	GATA Binding Protein 3
GWAS	Genome-wide association studies
HSCs	Hematopoietic Stem Cells
H3K4me3	Histone Subunit 3 Lysine 4 Trimethylation
H3K27ac	Histone Subunit 3 Lysine 27 Acetylation
ITH	Intra-tumoural Heterogeneity
MSRE	Methylation-Sensitive Restriction Endonuclease
MuTE	Mutation of *TAL1* Enhancer
MYB	MYB Proto-Oncogene, Transcription Factor
RUNX1	Runt-related Transcription Factor 1
SNP	Single-nucleotide polymorphism
SNV	Single-nucleotide variant
TAL1	T-cell Acute Lymphocytic Leukaemia
T-ALL	T-cell acute lymphoblastic leukaemia
TF	Transcription Factor

Abstract

T-cell acute lymphoblastic leukaemia (T-ALL) is the hyperproliferative transformation of T-cell lymphoid progenitor cells within the blood and bone marrow and is extremely heterogeneous. T-ALL has been linked to the overexpression of transcription factors, such as TAL1, that is specific within the late-cortical subtype of T-ALL. This project has utilised clonal cell line populations for testing phenotypic intra-tumoural heterogeneity seen within cancer cell lines, such as the Jurkat cell line, to generate clonal populations relative to the parental cell line they are derived from at passages 1, 5 and 9 as a cost-effective model. We tested the phenotype of proliferation using a carboxyfluorescein succinimidyl ester (CFSE) assay which identified Jurkat clonal populations as highly proliferative and displayed lower expression of the *TAL1* gene, relative to the parental cell line using real-time PCR analysis. We also identified four differentially bound putative regulatory element sites using bioinformatics analysis of publicly available data. This analysis displayed a Jurkat-specific predicted intragenic regulatory element and intergenic enhancer regions that map to the known upstream *TAL1* Jurkat super-enhancer as stated by Mansour et al. (2014). DNA methylation is known to fine-tune intragenic and intergenic enhancer-mediated transcription. Thus, we used a methylation-sensitive restriction endonuclease (MSRE) assay that provided insight of dynamic and stable DNA methylation patterns at the intragenic and intergenic sites across the *TAL1* locus between Jurkat clonal populations, respectively, at passages 1 and 9. Finally, using MinION nanopore sequencing, we identified single-nucleotide variants common between Jurkat clonal populations tested at passages 1 and 9, which map to regulatory elements, SNPs in linkage disequilibrium across the *TAL1* locus and sites of predicted transcription factor binding, therefore suggesting regulatory functionality of these SNVs in the context of the TAL1-mediated T-ALL.

Chapter 1 - Background

1.1. Cell Differentiation Programs can be Hijacked by Cancer

Inter- and intra-cell signalling determines cell-lineage identity and is required for the coordinated function of cells and tissues within the body (Porcher et al., 2017). These coordinated events are regulated through cell- and tissue-specific programming of genes as well as the timing, magnitude and duration of gene expression. The production of hematopoietic stem cells (HSCs), which are precursors of myeloid and lymphoid lineages, and the processes of differentiation of HSCs, allow for the functioning of the innate and adaptive immune system in a coordinated fashion. Dysregulation of the differentiation of HSCs and/or the lymphoid and myeloid precursors can result in atypical cell proliferation and differentiation, ultimately leading to malignancy (Porcher et al., 2017).

1.2. Acute Lymphoblastic Leukaemia (ALL)

Acute lymphoblastic leukaemia (ALL) is the transformation and hyperproliferation of clonal lymphoid-progenitor cells in the blood and bone marrow of patients. This malignancy of B- or T- lymphocytes is the second most prevalent leukaemia in adults, with 300 cases diagnosed each year amongst children and adults in Australia (Chiaretti et al., 2016; Leukaemia Foundation, 2019). However, the incidence of ALL is bimodally distributed between children and adults, with the occurrence of the disease appearing between ages 2-4 and on average 50 and older (respectively) (Chiaretti et al., 2016). ALL develops in response to endogenous factors, such as genetic susceptibility, although the significant phenotypic and genetic differences between ALL subtypes poses challenges when identifying causative events and genetic changes that are the foundation of ALL development (Chiaretti et al., 2016).

The most common form of ALL is B-cell acute lymphoblastic leukaemia (B-ALL), and therapies for this subtype have greatly improved the survival rate to where 90% of children and 40% of adult patients with the disease remain in remission over the span of 5 years (Malouf and Ottersbach, 2018). The T-lymphoid progenitor equivalent (T-ALL), has a similar long-term survival rate within children, where over 85% of childhood cases undergoing therapy will be in remission for five years, however has a significantly higher rate of relapse with a median point of 1.2 years in contrast to 2.5 years in B-ALL seen amongst 125 T-ALL patients after cessation of treatment (Goldberg et al., 2003; Terwilliger and Abdul-Hay, 2017). Therefore, understanding of the genetic basis of T-ALL may provide insight into the higher rate of relapse seen for this subtype of ALL.

The development of T-ALL can be due to inherited genetic mutation(s) or sporadic somatic mutations (Girardi et al., 2017; Inaba et al., 2013). In the last decade, advancements in understanding the genetic basis of ALL has distinguished multiple subtypes within B-cell and T-cell ALL (Iacobucci and Mullighan, 2017). Both these subtypes of ALL are distinguished individually by unique structural DNA rearrangements and sequence mutations that affect critical steps in lymphoid development, cytokine receptor signalling, Ras-kinase signalling and tumour suppressive chromatin remodelling processes (Iacobucci and Mullighan, 2017). Analysis of the T-cell subtype of ALL shows that 50% of patients harbour chromosomal translocations involving T-cell receptor genes (TCRA-TCRD and TCRB) and genetic alterations to pivotal transcription factor (TF) genes (Table 1.1) (Iacobucci and Mullighan, 2017).

1.3. A Subtype of ALL: T-Cell Acute Lymphoblastic Leukaemia (T-ALL)

T-ALL is a highly malignant subtype of ALL that accounts for approximately 20% of ALL cases amongst children and adults (Litzow and Ferrando, 2015). T-ALL is highly

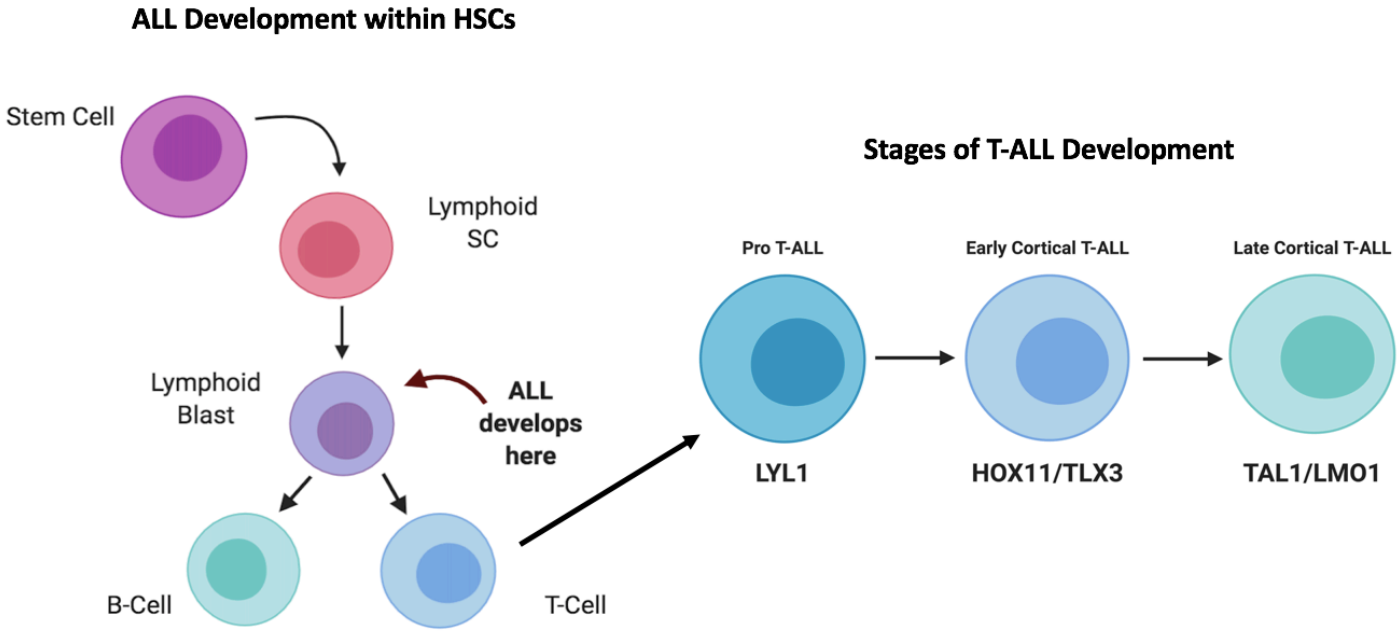
| Table 1.1. Summary of gene categories and targets that have genetic rearrangements or mutations that are commonly found within T-cell acute lymphoblastic leukaemia (T-ALL) [1].

			Frequency (%)	
Gene Category	**Gene Target**	**Genetic Rearrangement**	**Paediatric**	**Adult**
Basic Helix-Loop-Helix (bHLH)	TAL1	t(1;14)(p32;q11) t(1;7)(p32;q34) 1p32 deletion	30	34
	TAL2	t(1;7)(q34;q32)	1*	
	LYL1	t(7;19)(q34;q13)	1*	
	BHLHB1	t(14;21)(q11.2;q22)	1*	
LIM-only domain (LMO)	LMO1	t(11;14)(p15;q11)	6*	
	LMO2	t(11;14)(p13;q11)	10	2
HOXA Homeobox (HOX)	TLX1/HOX11	t(11;14)(p15;q11)	8	20
	TLX3/HOX11L2	t(11;14)(p15;q11)	19	9
	NKX2.1	inv(14)(q11.2q13)	8*	
	NKX2.2	t(14;20)(q11;p11)		
NOTCH1-related	NOTCH1	t(7;9)(q34;q34.3)	50	57
Cell Cycle Alterations	CDKN2A/2B	9p21 deletion methylation	61/58	55/46
	CCND2	T(7;12)(q34;p13)	1*	
Cell Growth Tumour Suppressors	RB1	13q14 deletion	12*	
	CDKN1B	12p13 deletion	2*	
	MYC	t(8;14)(q24;q11)	1*	
	WT1	Inactivating mutation/deletion	19	11
	LEF1	Inactivating mutation/deletion	10	2
	ETV6	Inactivating mutation/deletion	8	14
	BCL11B	Inactivating mutation/deletion	10*	
	RUNX1	Inactivating mutation/deletion	8	10
Signal Transduction	GATA3	Inactivating mutation/deletion	5	3
	PTEN	Inactivating mutation	19	11
	NRAS	Activating mutation	14	9
	NF1	Inactivating mutation/deletion	4	4
	JAK1	Activating mutation	5	7
Chromatin Remodelling	IL7R	Activating mutation	10	12
	EZH2	Inactivating mutation/deletion	12	12
	SUZ12	Inactivating mutation/deletion	11	5
	EED	Inactivating mutation/deletion	5	5
Proto-oncogene	MYB	t(6;7)(q23;q32)	7	17

[1]Genetic alterations seen within the genes listed are seen within >2% of T-ALL cases. This table displays a compilation of

genetic alterations and their frequencies based on the studies by Girardi et al. (2017) and Van Vlierberghe and Ferrando (2012). * Frequencies for pediatric and adult cases could not be separated.
[2]Yellow, green and orange boxes indicate characteristic gene mutations found within the subtypes of late cortical T-ALL, pro T-ALL and early cortical T-ALL, respectively. *LMO1* and *LMO2* are seen in some cases of late cortical T-ALL (Girardi et al. 2017).

heterogeneous with several subtypes that are distinguished by differences in the rate of survival and clinical prognosis, such as pro-T-ALL and cortical T-ALL (early and late) (Fig 1.1).



| Figure 1.1. T-Cell Acute lymphoblastic leukaemia (T-ALL) development occurs within the lymphoid differentiation process for T-cells.

T-ALL develops within the bone marrow at the early pro- and pre-T stages of T-cell development. However, T-ALL can progress to stages of maturation in the cortex and medulla within the thymus where T-cell differentiation is blocked (Khera et al., 2014). The subtypes of T-ALL are defined based on the maturation of the T-ALL cell through stages of pro-, pre-, early cortical, late cortical and medullary T-ALL. Key genes are dysregulated within T-ALL subtypes and this is linked to clinical prognosis. The subtype pro-T-ALL is characterised by the *LYL1* gene in which LYL1-specific T-ALL is known to have a poor prognosis. The subtypes of early cortical and late cortical T-ALL are characterised by TLX3/HOX11 and TAL1/LMO1 respectively. Clinical prognosis is better within the early cortical stages of T-ALL, relative to the late cortical stage of T-ALL (McCormack et al., 2013; You et al., 2015). Created with BioRender.

Due to subtypes of T-ALL harbouring different genetic rearrangements and smaller substitutions/insertions and deletions (indels), these subtypes can be distinguished and linked to different clinical outcomes (Yadav et al., 2016).

T-ALL is subdivided into further subtypes dependent on intra-thymic differentiation such as pro-T-ALL, pre-T-ALL, cortical T-ALL and medullary T-ALL (Litzow and Ferrando, 2015). Early and late stages of the cortical subtype of T-ALL display dysregulated expression of genes *TLX1* and *TAL1*, respectively, however prognosis is favourable within *TLX1*-specific T-ALL in comparison to *TAL1*-specific T-ALL. Pro-T-ALL is distinguished by the overexpression of the *LYL1* gene within early T-cell precursor cells and is linked to a poor prognosis similar to *TAL1*-T-ALL (McCormack et al., 2013; You et al., 2015) (Fig 1.1). These gene expression profiles have shown that T-ALL cases can be subtyped into mutually exclusive groups, based on the expression of these key oncogenes (Tan et al., 2019) (Table 1.1). Therefore, linking genes to specific subtypes in T-ALL can be used to understand the impact the genetic mutations within these genes have on molecular pathways that induce differential clinical outcomes.

An example of a chromosomal translocation that induces oncogenic activity within T-ALL is the, t(7;9)(q34;q34) translocation between the T-cell receptor B (*TCRB*) gene and *NOTCH1,* which is a common mutation seen within 60% of T-ALL pathogenesis (Litzow and Ferrando, 2015) (Table 1.1). This translocation fuses the *NOTCH1* gene to regulatory elements, such as promoters and enhancers, of the *TCRB* gene and dysregulates the expression of NOTCH1 (Yamamoto et al., 2013). Other T-ALL subtypes commonly show T-cell receptor gene chromosomal translocations (such as the *TCRB*/*NOTCH1* translocation) that result in the atypical expression of TF oncogenes: *LMO1*, *LMO2*/*LYL1*, *TLX1* and *TLX3* for late cortical, early pro- and early cortical T-ALL subtypes, respectively (Table 1.1 – green and orange boxes). In

addition to T-cell receptor gene translocations, genetic translocations and mutations can also confer loss of TF activity such as differentiation, tumour suppressor and cell-cycle inhibitor TFs (e.g. TAL1, RUNX1, ETV6, EZH2, CDKN2A and CDKN1B) that become dysregulated within T-ALL and confer malignancy, specifically seen in *TAL1* overexpression within the late cortical subtype of T-ALL (Table 1.1 – Yellow box) (Litzow and Ferrando, 2015; Van Vlierberghe and Ferrando, 2012).

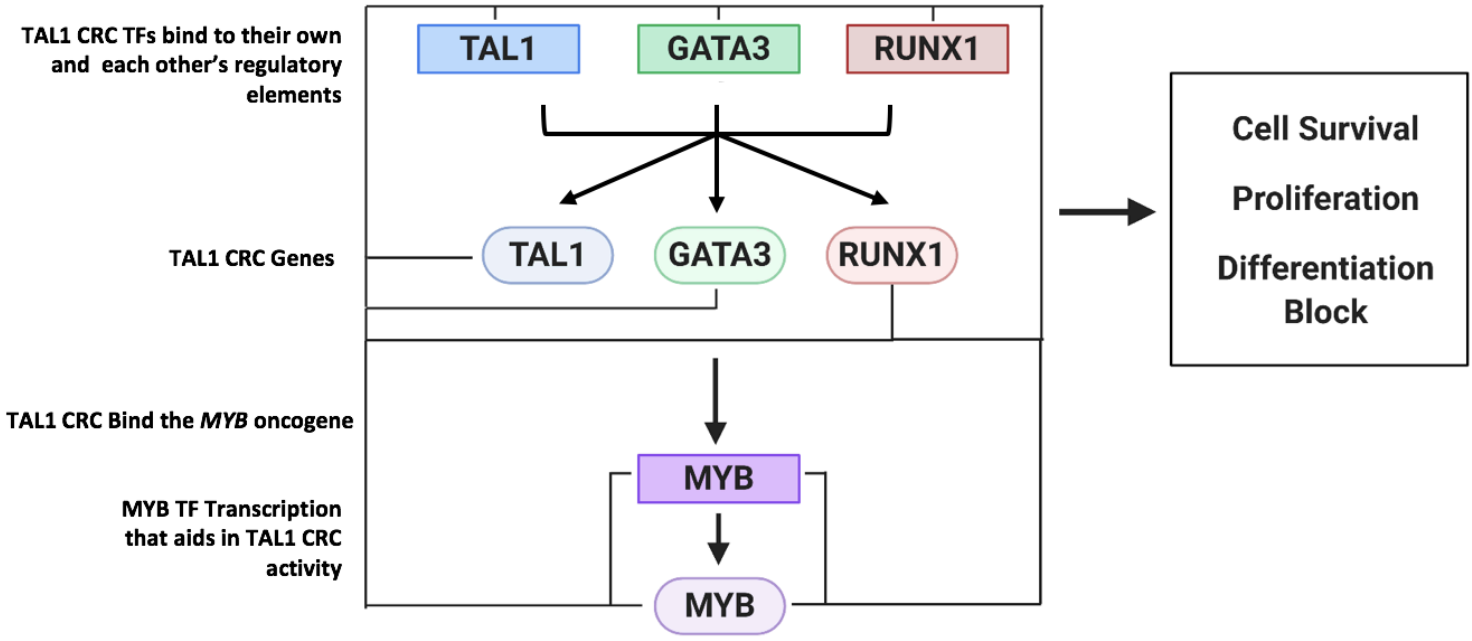
Chromosomal aberrations in T-ALL are a common occurrence, specifically involving translocations of the basic helix-loop-helix (bHLH), cysteine-rich (LIM-domain) and/or homeodomain genes (Sanda and Leong, 2017). Currently, T-ALL patients with activating oncogenic mutations for bHLH proteins, like *TAL1*, have a poor response to current therapeutics and are associated with high-risk failure (50% patient remission within five years) (Sanda and Leong, 2017). These include genes such as *MYC*, *TAL1*, *TAL2*, *LYL1* and *bHLHB1* in the bHLH category, *LMO1* and *LMO2* (LIM-domain genes) and *HOX11*/*TLX1* and *HOX11L2*/*TLX3* (Litt et al., 2013). These TFs regulate distinct differentiation pathways in T-cells, however are seen to be aberrantly co-expressed within their respective T-ALL subtypes (Sanda and Leong, 2017).

1.3.1 The *TAL1* gene in T-ALL

The T-cell acute lymphocytic leukaemia (*TAL1)* gene is located on chromosome 1p32 and encodes for a class I bHLH transcription factor TAL1 (Litt et al., 2013). The ectopic expression of TAL1 is specifically associated with the most common T-ALL subtype, late cortical T-ALL, accounting for 40-60% of all cases (Tan et al., 2019). The expression of *TAL1* is essential for the genesis of HSCs in embryogenesis and in turn, erythroid and lymphoid maturation (Litt et al., 2013). Within murine lymphoid progenitor cells, the expression of *TAL1* is silenced at the double negative ($CD4^-/CD8^-$) phase for thymic T-lymphocytes (Tan et al., 2019). However, the mutations already described within *TAL1* (Table 1.1) drive its continued expression, resulting in a differentiation block for developing T-cells (Sanda and Leong, 2017; Tan et al., 2019).

The translocation t(1;14)(p32;q11) is found in 3% of all T-ALL childhood cases and results in the over-expression of TAL1 through a fusion with regulatory elements within the *TCRA/D* oncogene (Van Vlierberghe and Ferrando, 2012). A more frequent *TAL1* translocation in 12-25% of late cortical T-ALL cases is a 90-kb deletion that fuses the first exon of the SCL interrupting locus (*SIL)* gene to the *TAL1* gene, increasing expression of TAL1 through the *SIL* regulatory elements (Carlotti et al., 2002; Janssen et al., 1993; Litt et al., 2013; Tan et al., 2019). Although these examples have identified translocations that increase *TAL1* expression, 60% of late cortical T-ALL cases show overexpression of *TAL1* in the absence of any detectable gene rearrangements within the *TAL1* gene (Litt et al., 2013). In the absence of large chromosomal rearrangements, smaller mutations may drive changes in *TAL1* expression. An example of this is the heterozygote 12-bp insertion described by Mansour et al. (2014), that creates a powerful enhancer that turns on *TAL1* gene expression. This suggests there may be other genetic changes associated with the *TAL1* locus that target regulatory elements and increase TAL1 expression.

The over-expression of TAL1 promotes T-ALL by two mechanisms. In the first mechanism, TAL1 forms a heterodimer with class A bHLH proteins such as HEB, E2A and E47, and also forms large complexes with TFs such as GATA1 and other LIM-only proteins (O'Neil et al., 2004). The TAL1 heterodimer recognises an E-box sequence, CANNTG, at promoter sites of targeted genes (Palomero et al., 2006). The overexpression of TAL1 inhibits the formation of E2A/HEB heterodimers and instead favours the formation of TAL1/E2A heterodimers. The E2A/HEB heterodimer aids in tumour suppressive functions through its involvement in the chromatin-remodelling complex, specifically inducing histone-deacetylase 1 and 2 activity (HDAC1 and HDAC2) (Litt et al., 2013). However, the formation of TAL1/E2A heterodimers promotes oncogenic transformation by interfering with E2A/HEB heterodimerisation (Sanda and Leong, 2017). It is this shift in heterodimer composition through TAL1



| Figure 1.2. TAL1 core regulatory circuit (CRC) positive auto-regulatory loop.

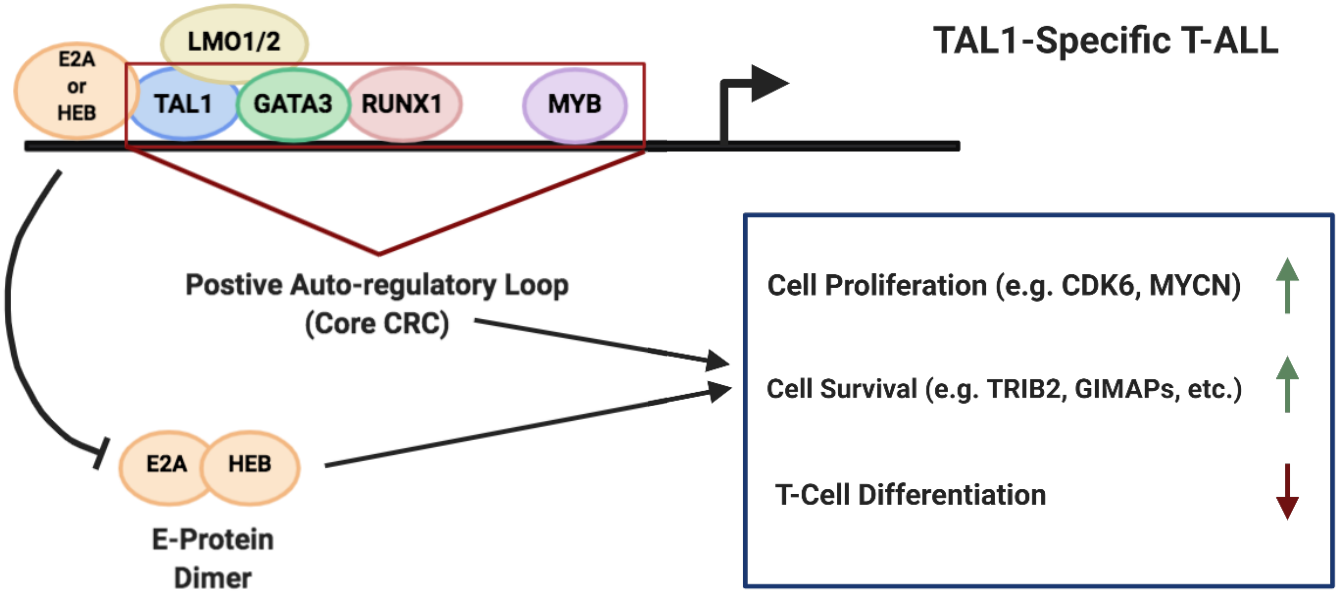
The TAL1 protein forms a complex with transcription factors which coordinate the regulation of downstream target genes for T-ALL malignancy, such as *MYB*. This autoregulatory loop of T-ALL core regulatory circuit (CRC) TFs (boxes), TAL1, GATA3 and RUNX1 bind to and initiate transcription of each other's and their own regulatory elements (circles), as well as the oncogene, *MYB* to maintain the T-ALL pathogenesis through mechanisms of cell survival, proliferation and blocking differentiation of the T-Cell through TRIB2, GIMAP protein, CDK6 and MYCN and blocking of E-protein heterodimerisation pathways (Sanda and Leong, 2017; Tan et al., 2019). Created with BioRender.

overexpression that disrupts E2A/HEB-dependent gene expression and switches to a T-ALL gene expression program (Sanda and Leong, 2017; Sanda et al., 2012).

In the second mechanism, the dysregulated expression of TAL1 allows continued TAL1-dependent gene expression in late cortical T-ALL cells through combining with TFs, GATA3 and RUNX1, that are also expressed in T-ALL (Table 1.1). TAL1 binds to similar regulatory sites within genes with GATA3, and RUNX1 (Sanda et al., 2012). Although TAL1, GATA3, and RUNX1 do not directly interact, they cooperate to form

a core TF regulatory circuit (CRC) that binds to regulatory elements within genes that sustain the T-ALL pattern of gene expression as well as co-occupying their own and each other's regulatory elements (Sanda et al., 2012). This creates an autoregulatory loop unique to T-ALL pathogenesis in late cortical stages (Sanda and Leong, 2017) (Fig 1.2). Hence, the TAL1 CRC reinforces and promotes the stability of the T-ALL leukemogenicity program (Litt et al., 2013).

The feed-forward loops established by TAL1 creates a process of 'forced' expression of TFs which are part of a set of 'master TFs'. These master TFs then dysregulate gene expression programs that lead to a change in cell identity associated with T-ALL (Sanda and Leong, 2017). Chromatin-immunoprecipitation sequencing (ChIP-seq) experiments and gene expression assays confirm genome-wide occupancy of TAL1 and its CRC partners GATA3 and RUNX1 in T-ALL cells. ChIP-seq also shows co-occupancy of TAL1, GATA3 and RUNX1 at regulatory elements in key T-ALL-specific genes, including the oncogene *MYB* (Palomero et al., 2006). Independent studies show co-occupancy of regulatory DNA sequences by these TFs as seen at the *eR1* (*RUNX1* enhancer) and the *MuTE* region (Mutation of *TAL1* enhancer) that introduce MYB-binding motifs within late cortical T-ALL (Mansour et al., 2014; Tan et al., 2019; Yamamura et al., 2017). The *eR1* enhancer element resides between the promoters of the *RUNX1* gene and drives the expression of *RUNX1* within HSCs, but is also seen within other tissue types (Yamamura et al., 2017). The *MuTE* region mutation is an insertion of 12-bp that introduces binding motifs for the oncogenic TF, MYB, and is linked to the overexpression of the *TAL1* gene downstream (Mansour et al., 2014). These examples support the idea that the TAL1 CRC maintains a program of gene expression required for T-ALL leukemogenesis. In summary, TAL1-dependent dysregulation increases the CRC activity, increasing the expression of genes, including TAL1 itself, in the context of late cortical T-ALL (Fig 1.3).



| Figure 1.3. Model of TAL1-induced transformation in late cortical T-ALL.

TAL1 is normally not expressed during maturation stages of T-cell development, however due to chromosomal and intrachromosomal rearrangements or mutations, changes in regulatory elements such as enhancer elements may occur that induce TAL1 overexpression. The two TAL1-dependent mechanisms of leukemogenesis are shown: assembly of the T-ALL CRC initiation with GATA3 and RUNX1 and TAL1 heterodimerisation of either E2A or HEB, leading to the inhibition of E2A/HEB heterodimerisation for tumour suppressive functions. This leads to aberrant cell proliferation, cell survival and a differentiation block by affecting downstream targets, such as *CDK6* and *MYCN*, *TRIB2* and *GIMAP* protein genes and additional generic alterations in key T-ALL genes such as *NOTCH1* (Sanda and Leong, 2017; Tan et al., 2019). Created with BioRender.

1.4. Epigenetic Landscapes: DNA Methylation

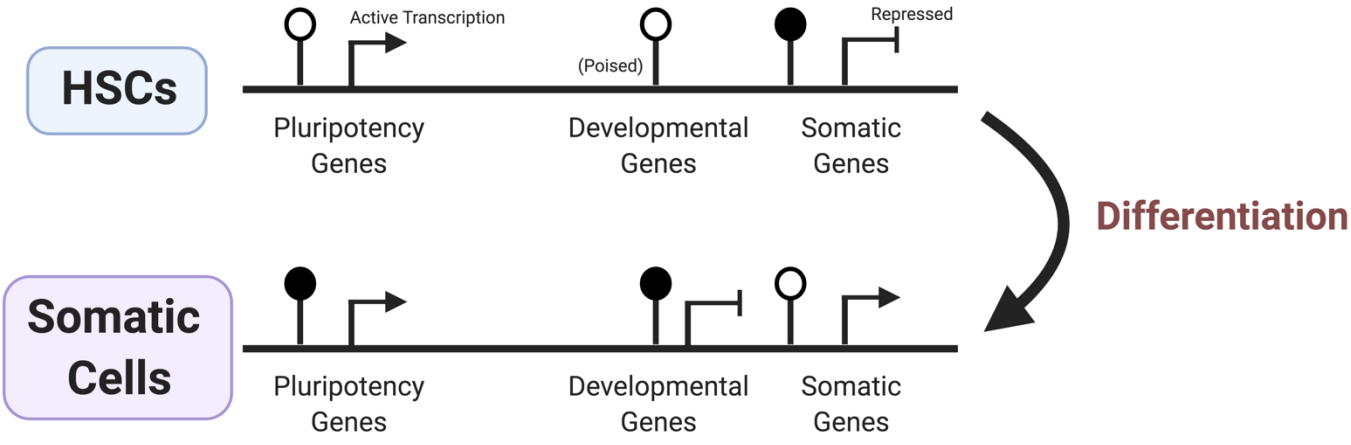
As already discussed, a number of large chromosomal rearrangements leading to over-expression of *TAL1* have been found in T-ALL (Table 1.1). However, in 60% of late cortical T-ALL cases showing an over-expression of TAL1, there is no evidence of such chromosomal rearrangements. The identification of a small insertion mutation upstream of the *TAL1* gene and the resulting increase in *TAL1* expression suggests

there may be other as-of-yet uncharacterised small mutations and/or epigenetic modifications within regulatory DNA sequences at the *TAL1* locus that result in increased expression of *TAL1* in T-ALL.

An epigenetic modification is defined as a reversible and stable change in gene expression without the alteration of a DNA sequence (Luczak and Jagodziński, 2006). These modifications include DNA methylation and histone subunit post-translational modifications such as methylation and acetylation. Alteration in DNA methylation is associated with a variety of haematological disorders, and DNA methylation can be used as a prognostic marker due to the significant impact it has on gene expression (Luczak and Jagodziński, 2006). Epigenetic alterations, such as DNA methylation, show specific signatures in cancers which can be used to further classify malignancies, such as T-ALL, and establish a connection between epigenetic profiles and pathogenesis (Haider et al., 2019).

DNA methylation consists of the addition of a methyl group to a cytosine base within CpG dinucleotides by DNA methyltransferases (Kulis and Esteller, 2010). The presence of a methyl group bound to a CpG dinucleotide promotes condensation of chromatin and transcriptional silencing that can be tissue-specific and is also vital for stabilisation of chromosomes (Luczak and Jagodziński, 2006) (Fig 1.4, Differentiation-Specific DNA Methylation). In non-cancer cells, it allows for the regulation of gene expression by initiating and maintaining stable gene silencing (Kulis and Esteller, 2010). With its association with post-translational histone modifications, it is crucial for the architecture of chromatin and its effect on transcriptional activity (Kulis and Esteller, 2010). For example, proteins such as m^5CpG-binding domain (MBD) and m^5CpG-binding-proteins (MeCP) induce DNA methylation-dependent repression by interacting with HDAC1 and HDAC2 to alter patterns of chromatin remodelling (Luczak and Jagodziński, 2006).

CpG islands are defined as stretches of DNA 500-1500bp long with a CG:GC ratio of more than 0.6 (Bird et al., 1995). These CpG islands are typically found at promoters and contain the 5' end of the transcript (Bird et al., 1995). These islands are frequently found at transcription start sites (TSSs) which are unmethylated within healthy cells (Qu et al., 2012).



| Figure 1.4. DNA methylation patterns amongst haematopoietic stem cells (HSCs) and the changes of DNA methylation during differentiation within somatic cells.

Promoter methylation is indicated by filled in markers and unmethylated promoters are indicated by the open markers. HSCs typically have patterns of DNA methylation that allow for the transcription of pluripotency and developmental genes during development, whereas specific somatic differentiation genes are inhibited by DNA methylation of associated promoters. However, within somatic cells that are differentiated, genes involved in pluripotency and developmental pathways are inhibited through DNA methylation in conjunction with somatic differentiation gene expression, allow for the differentiation of somatic cells. This ensures that the development of differentiating cells is mutually exclusive when controlling gene expression from pluripotency/development genes and somatic gene expression. Created with BioRender.

This contrasts with the remaining regions of the genome, which are predominantly methylated (King et al., 2016). Hypomethylated regions are not limited to TSSs but include intragenic and intergenic regions that are associated with non-coding

regulatory DNA elements such as enhancers (Qu et al., 2012). Intragenic regions are defined as sequences within a gene body (i.e. defined by exons and intervening introns), whereas intergenic regions are the regions of the genome outside of a gene body. Therefore, DNA methylation is found throughout the genome, with dynamic changes in DNA methylation at TSSs, intergenic and intragenic regions and is associated with changes in cell identity and state (Varley et al., 2013). In the context of cancer, CpG islands can become atypically hypermethylated, leading to the suppression of tumour suppressor genes and their associated pathways (Kraszewska et al., 2012).

Within T-ALL, DNA methylation has prognostic relevance that is based on the CpG island methylation phenotype (Haider et al., 2019). Within the *TAL1* gene, promoter hypomethylation is associated with dysregulated transcriptional control of the gene, leading to *TAL1*-overexpression and a worsened prognosis (Haider et al., 2018). However, changes in DNA methylation patterns are not limited to TSSs and promoter regions, therefore, a deeper understanding of the patterns of DNA methylation across intragenic and intergenic regulatory non-coding sequences within the *TAL1* locus can further identify how DNA methylation can influence gene expression in late cortical T-ALL.

1.5. Epigenetic Landscapes: Regulatory Elements - Enhancers

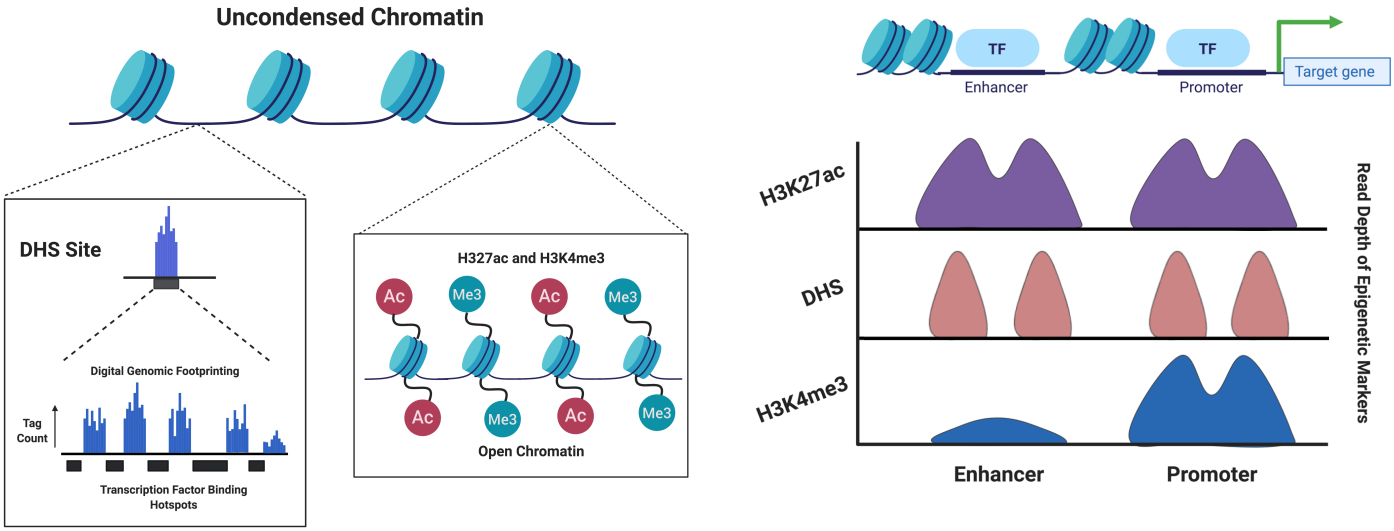
Our understanding of the relationship between DNA methylation and post-translational histone subunit modifications, such as acetylation and methylation, and their importance in maintaining inducible and cell-specific patterns of gene expression is in large part due to the Encyclopedia of DNA elements (ENCODE) project (Consortium, 2004; Davis et al., 2018; Qu and Fang, 2013; Sloan et al., 2016; The ENCODE Project Consortium, 2011). The ENCODE project has provided a map of the

epigenetic landscape of the human genome, which identifies putative transcription regulatory DNA elements through the co-localisation of specific patterns of DNA methylation and histone subunit methylation and acetylation (Davis et al., 2018). The ENCODE project is fundamental to understanding the 98% of the human genome that does not encode proteins, with early publications from the ENCODE project claiming 80% of the human genome can be assigned a "biochemical function" (The ENCODE Project Consortium, 2011). This knowledge is in stark contrast to the historical view of non-coding DNA as 'junk DNA' (Palazzo and Gregory, 2014). It is with this understanding that key regulatory elements such as enhancers and promoters are mapped across the genome.

Promoters are proximal regulatory regions near TSSs of genes and mediate the binding of RNA polymerase II through promoter recognition of TFs (Carelli et al., 2018). Most promoters are CpG-rich, whilst other regulatory regions, such as enhancers are CpG poor (Carelli et al., 2018). However, it is the accessibility of chromatin at these regions that determines their functionality in the context of the spatial and temporal regulation of gene transcription (The ENCODE Project Consortium, 2011). Through ENCODE and other studies on putative enhancers, as many as 400,000 distinct enhancer elements are mapped in specific human cell lines (Calo and Wysocka, 2013). Enhancers can interact with TSSs and promoters independently of location (up to 1 million base pairs away) (Tippens et al., 2018). Currently, the mechanism of promoter-enhancer interactions is proposed to involve "looping". Looping interactions involve enhancer-associated factors and co-factors that directly interact with the promoter through the looping of the intervening DNA sequence (Calo and Wysocka, 2013). This mechanism has been validated through the results obtained using chromosome conformation capture technologies such as 3C and derivatives 4C, 5C and Hi-C (Calo and Wysocka, 2013).

The core of an enhancer (200-500bp) is recognised by the presence of clustered recognition sites for multiple TFs. This forms a platform for non-DNA binding transcription coactivators such as mediator complexes, CREB-binding proteins and p300 (Jia et al., 2019). The combinatorial mechanisms of these activating factors ensures the integration of intrinsic and extrinsic cues within the cellular environment (Witte et al., 2015). Flanking sites of these TF binding regions are flagged with specific histone subunit modifications that are used to identify enhancer activity at specific loci (Benveniste et al., 2014). Active enhancers are associated with high levels of histone subunit 3 lysine 27 acetylation (H3K27ac) and low levels of histone subunit 3 lysine 4 trimethylation (H3K4me3) (Jia et al., 2019) (Fig 1.5). Another signal associated with active enhancers is the occupancy of TFs at nucleosomal depleted sites (Teif et al., 2017). These sites display high sensitivity to cleavage by the enzyme deoxyribonuclease I (DNaseI), that cleaves DNA. With the combination of DHS and chromatin modification data provided by ENCODE and other studies, a robust identification method for active enhancer sites in the genome can be utilised to understand gene regulatory landscapes (Teif et al., 2017) (Fig 1.5).

Enhancers can be located away from a target gene locus (intergenic) or embedded within the gene body in an intron (intragenic). Intergenic enhancers can be seen as the "classical" type of enhancer that influences transcriptional activity of genes in nearby loci by looping and contacting a promoter. In the case of intragenic enhancers, gene regulation may involve the "classical" function of looping and contact with adjacent genes, or it may involve a different mechanism of transcriptional regulation of the gene in which it is embedded. This alternate mechanism involves the RNA polymerase II-mediated transcription of short lived enhancer RNAs (eRNAs) (Cinghu et al., 2017) (Fig 1.6A).

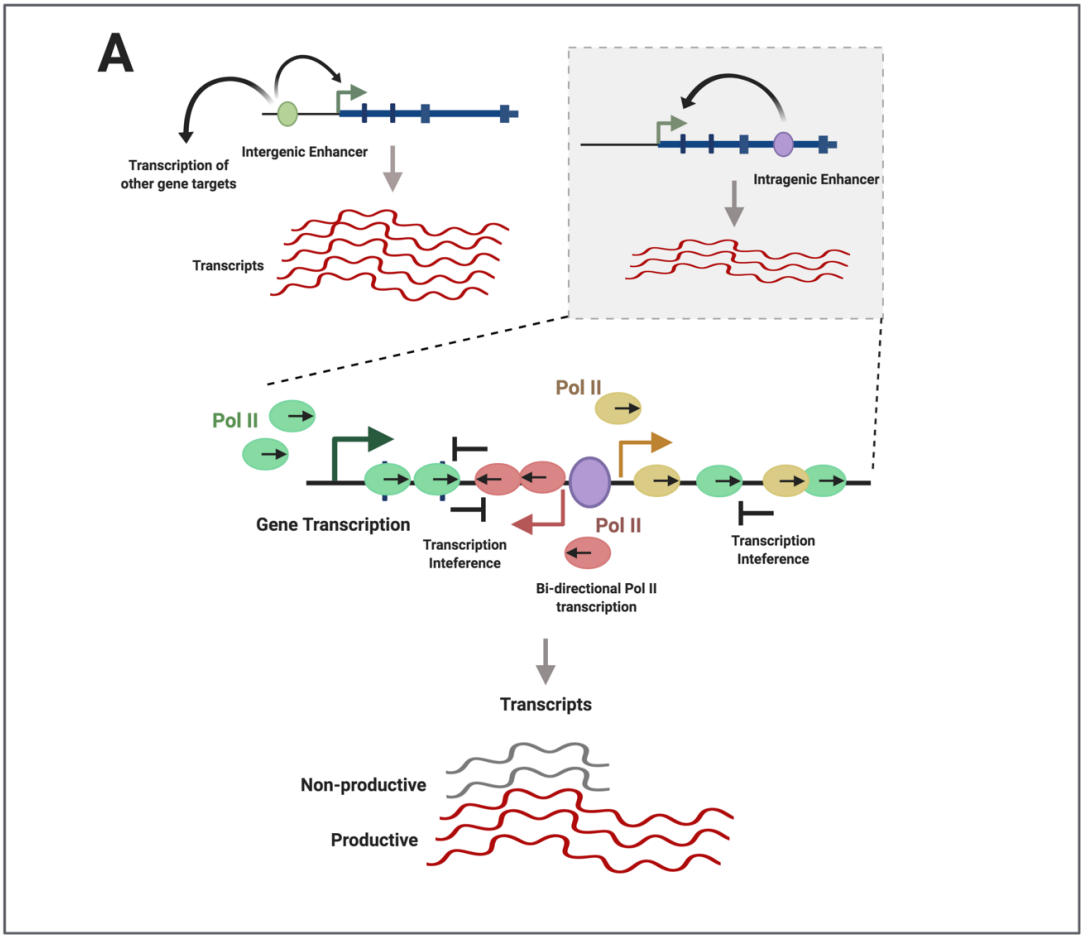


| Figure 1.5. Open chromatin is identified by H3K27ac and H3K4me3 and DNAse1 hypersensitivity (DHS).

Left: Digital genomic footprinting creates a map of DNaseI cleavage at single base-pair resolution across the genome. TF binding at these sites is marked by decreased cleavage (tag counts), leaving a "footprint" flanked by increased cleavage by DNAseI. ***Middle:*** Open chromatin can also be marked and initiated by histone modifications such as H3K27ac and H3K4me3. ***Right:*** Enhancers are marked by high levels of H3K27ac and DHS, while promoters are differentiated from enhancers through the presence of H3K4me3 (Sullivan et al., 2015; Vierstra and Stamatoyannopoulos, 2016). Created with BioRender.

It is well known that transcriptionally active genes show hypomethylation at promoter sites and increased methylation throughout the body of a gene (intragenic regions) (King et al., 2016; Neri et al., 2017; Rauscher et al., 2015; Varley et al., 2013). It is now shown that gene body methylation regulates intragenic enhancers, whereby methylation within the gene body prevents cryptic transcription from intragenic enhancer sites. These sites would otherwise promote RNA polymerase II-dependent bi-directional transcription of short eRNAs, which are believed to inhibit RNA polymerase II transcription from the promoter (Fig 1.6B – right side) (Cinghu et al., 2017; King et al., 2016; Sharifi-Zarchi et al., 2017; Weigel et al., 2019). This shows the importance of intragenic enhancers in the attenuation or fine tuning of the transcription of the gene in which they are embedded (King et al., 2016). This model would also suggest that intragenic DNA methylation is dynamic and is regulated according to gene activating or repressing signals (Varley et al., 2013). In contrast, intergenic DNA hypermethylation is associated with downregulation of functionally-associated genes as seen in pre B-ALL (Almamun et al., 2017). This demonstrates that the correlation between DNA methylation and gene expression is related to the methylation status of intragenic and/or intergenic enhancers (Almamun et al., 2017) (Fig 1.6B). Consequently, when considering the possibility of genetic variants within the *TAL1* locus, it is also important to consider intragenic as well as intergenic enhancers.

Other enhancer types that are also seen within intragenic and intergenic regions of the genome are known as super-enhancers (Khan et al., 2018a). The super-enhancers regulate the expression of cell identity genes and a number of disease-associated genetic variants map to these super-enhancers (Khan et al., 2018a) (Fig 1.7). Super-enhancers are regions of clustered enhancers that are occupied by master TFs, chromatin modifiers and/or histone markers of active activity (Jia et al., 2019). Super-enhancers are more enriched for active chromatin markers and more transcriptionally


A
Transcription of other gene targets
Intergenic Enhancer
Transcripts
Intragenic Enhancer
Pol II
Pol II
Gene Transcription
Transcription Inteference
Pol II
Bi-directional Pol II transcription
Transcription Inteference
Transcripts
Non-productive
Productive

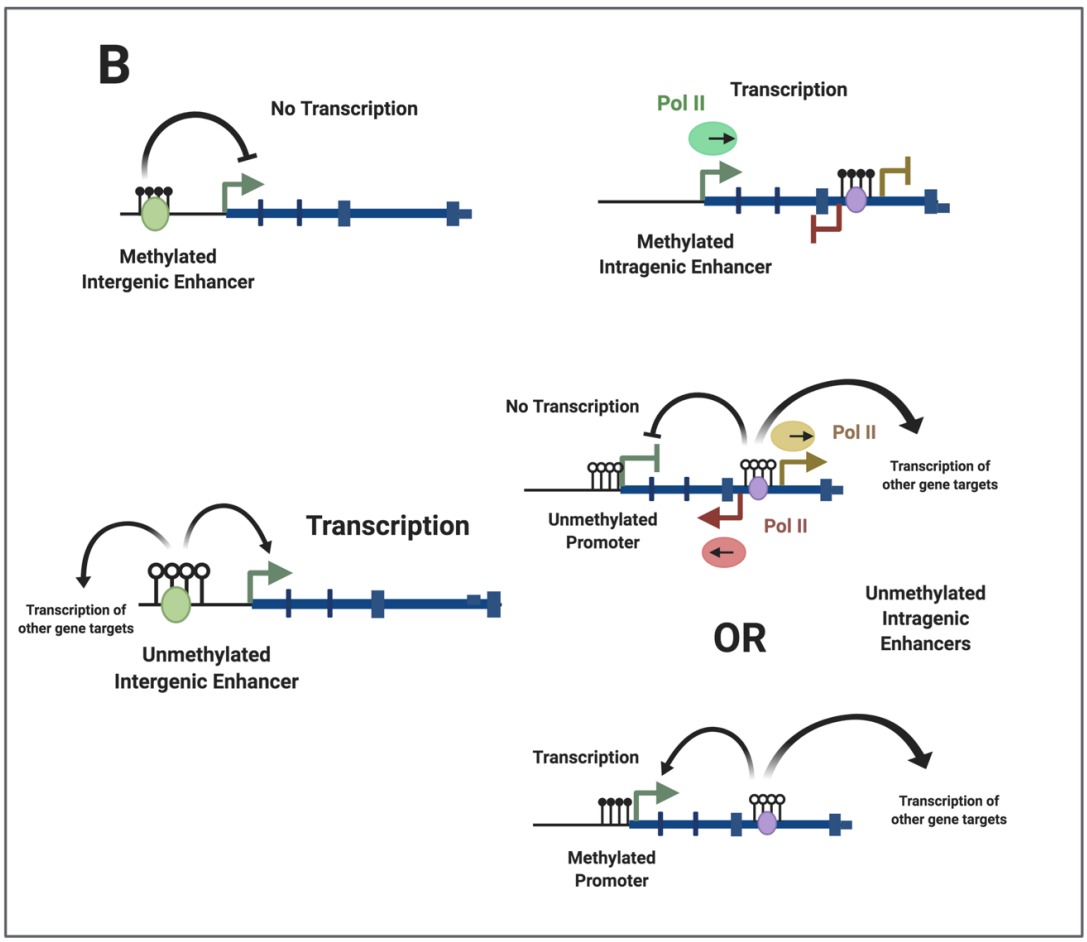

B
No Transcription
Methylated Intergenic Enhancer
Pol II
Transcription
Methylated Intragenic Enhancer
No Transcription
Pol II
Transcription of other gene targets
Unmethylated Promoter
Pol II
Transcription
Transcription of other gene targets
Unmethylated Intergenic Enhancer
OR
Unmethylated Intragenic Enhancers
Transcription
Transcription of other gene targets
Methylated Promoter

Figure 1.6 (A-B). Model for transcription at intergenic and intragenic enhancers and the influence of DNA methylation at intragenic and intergenic enhancer sites for transcription.

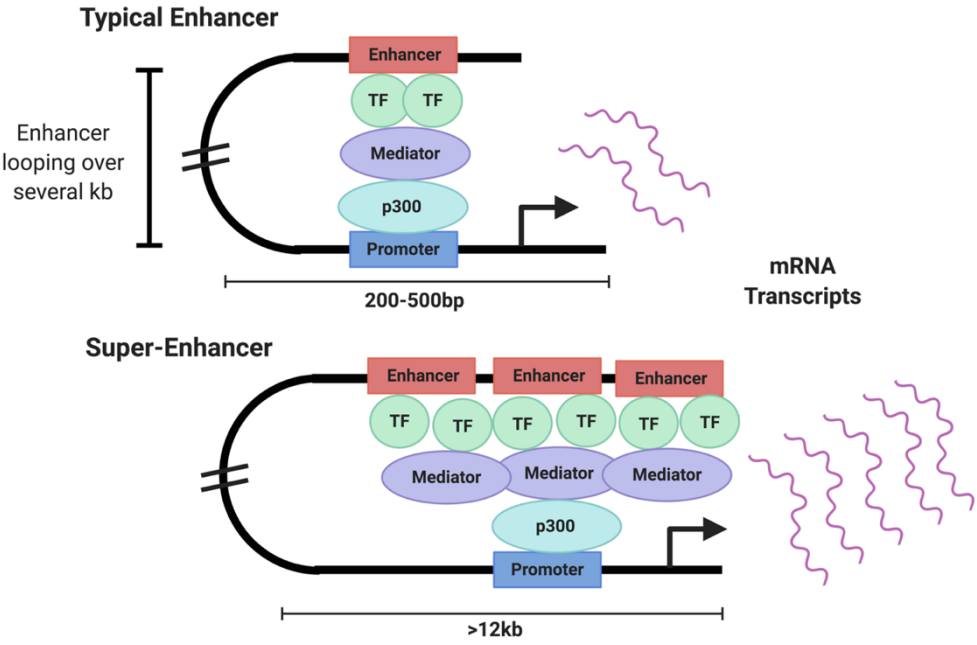
(Previous Page). A. ***Left:*** Intergenic enhancers loop and bind to promoter regions of genes to initiate transcription. ***Right:*** Intragenic enhancers inhibit the expression of the gene in which they are located. Bi-directional transcription from an intragenic enhancer (red and yellow circles) attenuates promoter-dependent transcription (green circle) during productive elongation through transcription interference with RNA Pol II. **B.** A model for DNA methylation regulation of enhancer activity. ***Left:*** DNA methylation at intergenic enhancers is associated with reduced levels of transcription of the target gene. When intergenic enhancers are unmethylated, functional enhancer-promoter contacts are formed, and transcription occurs at gene targets. ***Right:*** Consistent with the known inverse pattern of DNA methylation between the promoter and the gene body, it is proposed that methylated intragenic enhancers are transcriptionally inactive and do not inhibit RNA Pol II-dependent transcription from the promoter. A loss of methylation at the intragenic enhancer is thought to allow for attenuation or fine-tuning of promoter-dependent transcription from the gene in which it is embedded, and/or contact and transcription from promoter(s) of other gene(s). (Figure adapted from Cinghu et al. (2017) and Kowalczyk et al. (2012)). Created with BioRender.

active than typical enhancers specific for cell-lineage biological functions (Khan et al., 2018a) (Fig 1.7). It is currently unknown whether super-enhancers represent a small clustering of typical enhancers or act as an independent element with multiple facets of enhancer activity (Jia et al., 2019).

1.5.1. Enhancers within T-ALL – Jurkat Super Enhancer

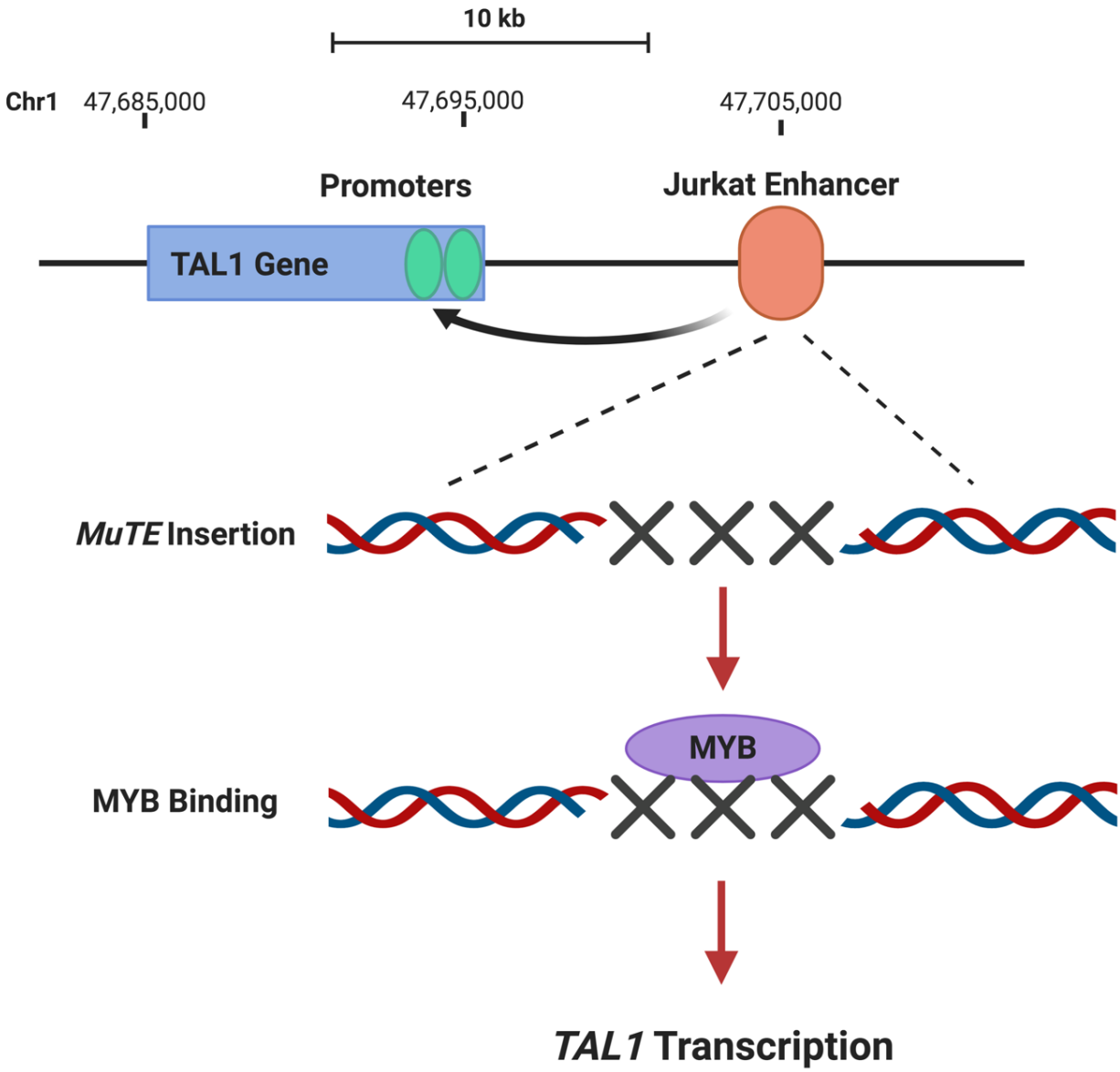
The translocation t(1;14)(p34;q11) between the *TCRB* and *TAL1* gene and SIL-TAL deletions are found in 2% and 20% of T-ALL cases respectively, yet more than half of the TAL1-specific cases lack *TAL1* lesions, therefore suggesting unidentified

epigenetic dysregulation patterns within TAL1-specific T-ALL (Navarro et al., 2015). Through ChIP-seq analysis, Jurkat cells are shown to have a strong H3K27ac signal upstream of the *TAL1* TSS that spreads towards the first exons of the gene (Mansour et al., 2014). A further study shows a high density of enrichment and breadth for H3K27ac spanning over -20kb to +10kb from the *TAL1* TSS (Rivera-Reyes et al., 2016). The *TAL1*-specific super-enhancer is enriched in the late cortical T-ALL Jurkat cell line relative to foetal thymocytes, hematopoietic (CD34+) stem and progenitor cells, and to other T-ALL cell lines such as DND41 and RPMI-8402 (Rivera-Reyes et al., 2016). The *TAL1* super-enhancer loops to make contact with the *TAL1* TSS, therefore playing a role in the overexpression of *TAL1* (Mansour et al., 2014).



| Figure 1.7. Differences between typical enhancers and super enhancers.

Typical enhancers range in a size of 200-500bp, and are bound by TFs, mediator proteins and the p300 complex when looping with a promoter to initiate transcription. Super-enhancers are clusters of enhancer modules that regulate the transcription of genes and are exponentially more potent in regulating gene expression relative to typical enhancers. Super-enhancers can bind multiple sets of TFs as well as mediator proteins to establish an increased function in regulating gene expression (Khan et al., 2018a). Made with BioRender.



| Figure 1.8. The organisation of *TAL1* loci within the Jurkat T-ALL cell line displaying an acquired super-enhancer upstream of the *TAL1* gene (chr1:47,705,000).

The acquired super-enhancer maps to a 12-bp heterozygote insertion mutation (mutation of TAL1 enhancer – *MuTE*). This insertion provides a binding site for the oncogenic MYB transcription factor, as well as other members of the TAL1 CRC. The Jurkat enhancer directly loops with the *TAL1* promoters (1b and 1a - green) to induce the overexpression of TAL1 (Mansour et al., 2014). Created with BioRender.

The *TAL1*-specific super-enhancer in the Jurkat T-ALL cell line is formed as a result of a heterozygous 12-bp insertion (GTTACCAAACGG) at the core of the super-enhancer site (Fig 1.8). The binding of TAL1 CRC TFs is seen in the TAL1-T-ALL Jurkat cell line at the *MuTE* insertion site (Mansour et al., 2014). Along this *MuTE* region, oncogenic master TF MYB binding motifs were identified *de novo*, completing the auto-regulatory loop of the TAL1 CRC (Sanda et al., 2012). CRISPR/Cas9 genomic editing to delete the *MuTE* insertion in Jurkat cells was unsuccessful, possibly reflecting the dependency of the Jurkat cell line on this region and *TAL1* expression for cellular proliferation (Rivera-Reyes et al., 2016).

1.7. Heterogeneity within Cancer Cell lines

The identification of a heterozygous somatic mutation upstream of *TAL1* is associated with monoallelic expression of *TAL1* in Jurkat T-cells (Mansour et al., 2014). However, the investigation of other key epigenetic modifications, such as DNA methylation, that could contribute to *TAL1* gene expression has not been thoroughly investigated in combination with mapped intragenic and intergenic enhancer regions within the *TAL1* locus.

It is evident that clonal heterogeneity within T-ALL patients could be a possible contributor to malignancy, as cells within a single tumour may be composed of multiple cell populations that may harbour different phenotypes and underlying genetics (Stockholm et al., 2007). Phenotypic heterogeneity is also seen within *in vitro* cell lines grown in controlled environments (Stockholm et al., 2007). These phenotypes between clonal cells can be diverse, as seen through examples of malignant cell lines in culture reverting to a stable non-malignant phenotype (Lavrovsky et al., 1992).

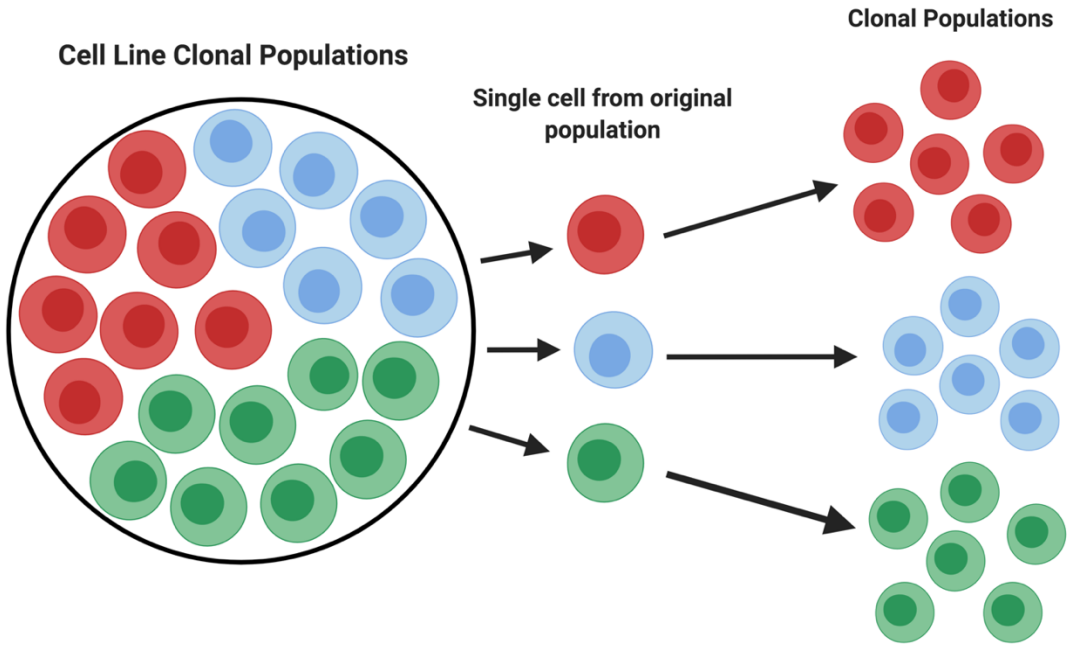
The frequent use of cell lines to investigate biological processes makes the understanding the heterogeneity of cell line phenotypes essential (Stockholm et al., 2007). Stockholm et al. (2007) propose a model that simulates the growth of clonal cell populations within cell lines and is a simple way to test phenotypic heterogeneity. It is found that 'intrinsic' spontaneous phenotypic switches occurred in the cell autonomously. This refers to the probability of phenotypic dynamics being based on underlying epigenetic and transcriptional regulatory mechanisms within each clonal population. This *in vitro* analysis can also contribute to a better understanding of *in vivo* processes as well as mimicking the heterogeneity established within a variety of cancer subtypes (Yadav et al., 2016).

Expanding from phenotypic heterogeneity, Martin-Pardillos et al. (2019) established a similar model of testing clonal cell lines at the phenotypic and molecular level using the breast cancer cell line, MDA-MB-231 (Martín-Pardillos et al., 2019). This was done to understand the accumulation of genetic alterations from an initial clonal population undergoing Darwinian selection, and complemented phenotype functional testing assays. Despite tumours being comprised of distinct clones that are phenotypically or genetically different, these clones interact to benefit one or more of the clones within the tumour, known as clonal cooperation (Neelakantan et al., 2015). Whilst competition amongst clones may result in dominant clones based on their survival fitness, the theory of clonal cooperation within populations demonstrates that complementary genetic alterations synergistically contribute to tumorigenesis and metastasis (Martín-Pardillos et al., 2019). This has been supported by single-cell sequencing studies, whereby clones cooperate to develop advantageous characteristics selective for atypical survival and angiogenesis (Martín-Pardillos et al., 2019). A conclusion of their study is that they have established a model in which the physical and chemical messenger interplay between cells in each population favours tumorigenic capacities of the breast cancer cell line.

The extent to which transcriptional heterogeneity contributes to phenotypic heterogeneity is still currently unknown (Ben-David et al., 2018). The study by Ben-David et al. (2018) expanded on ideas from the study by Martin-Pardillos et al. (2019) by using 106 human cell lines to analyse clonal diversity and the response of these clonal populations to various anti-cancer compounds. They demonstrated that ongoing instability within these cancerous cell lines can translate to heterogeneity, with each clone expressing differing responses to various drugs (Ben-David et al., 2018). Ben-David et al. (2018) further suggest that due to the instability of cancer cell lines, minimising variability within culture conditions between cell lines is recommended. This includes the generation of clonal populations derived from single cells, tracking passages and avoiding prolonged culturing of cell lines. Therefore, cancer cell lines must be thought of as highly heterogeneous due to pre-existing subclones within the population and from continuous instability while in culture (Ben-David et al., 2018) (Fig 1.9).

The Jurkat cell line has been used as an extensive model for T-cell activation and signalling, as well as a model for TAL1-specific T-ALL (Abraham and Weiss, 2004; Fernández-Ramos et al., 2017; Gioia et al., 2018; Mansour et al., 2014; Moharram et al., 2017). However, as research using the Jurkat cell line continues, irregularities concerning the Jurkat genetic and epigenetic profile have become apparent (Gioia et al., 2018). Using short-read sequencing, abnormalities of the Jurkat genome were identified in relation to damaging variants associated with cancer such as the apoptosis-related *BAX* gene (Gioia et al., 2018). Despite the documented instability of the Jurkat cell line, experimental designs used to minimise cell line variability within testing populations has not been commonly communicated, but can be tested to mimic intra-tumoural heterogeneity (ITH) and exploit cell line diversity as a cost-effective

method of profiling heterogeneity within cell line types (Fig 1.9) (Ben-David et al., 2018).



| Figure 1.9. Model of testing cancer cell line heterogeneity through the generation of clonal populations from single cells.

Cancer cell line populations are composed of subclones that may have genetic, transcriptional and phenotypic differences which may be due to epigenetic or genetic mutations or alterations that are selected for during tumorigenesis. A cell model to test this phenomenon of intra-tumoural heterogeneity (ITH) can be conducted through the isolation of single cells from the malignant parental (original) cell line population and grown to develop sub-clonal population cell lines to analyse for genetic, epigenetic and phenotypic differences to mimic processes seen within malignancy *in vivo*. Created with BioRender.

1.8 Hypothesis

This project will test the hypothesis that TAL1 is associated with the proliferation of late cortical T-ALL cells, and the isolation of clonal populations from the Jurkat cell line will show that differences in proliferation may be due to genetic or epigenetic differences at the *TAL1* locus between clones at varying passages of culture.

1.9. Objectives and Aims

Therefore, this project will be conducted with the following aims:

1. To generate clonal cell lines from a parental Jurkat cell line and to test these cell lines for differences in proliferation by using a cell-based carboxyfluorescein succinimidyl ester (CFSE) proliferation assay.
2. To test clonal cell lines for differences in the expression of T-ALL CRC genes, *TAL1, RUNX1, GATA3* and *MYB* using real-time PCR (qPCR) analysis.
3. Complete a bioinformatic analysis of the *TAL1* locus by using publicly available datasets from ENCODE and other genomic studies to map histone modifications, DNA methylation and DNaseI hypersensitivity sites in multiple different cell lines.
4. Use the bioinformatic analysis of the *TAL1* locus to target sites for DNA methylation analysis in the parental and clonal cell lines using a methylation-sensitive restriction endonuclease (MSRE) assay.
5. Clonal populations will be tested for genetic sequence differences through MinION nanopore sequencing of key regions across *TAL1*.

Chapter 2 - Characterisation of Jurkat Clonal Cell Lines

2.1. Introduction

The generation of clonal populations of cells is essential for minimising phenotypic, genetic and epigenetic variation in cancer cell lines *in vitro* (Ben-David et al., 2018; Martín-Pardillos et al., 2019; Stockholm et al., 2007). No single clone will exhibit all characteristics of a cancer, and it is the cooperation between genetically and phenotypically diverse clonal cell lines that drive cancer progression (Martín-Pardillos et al., 2019). Therefore, the study of clonal cell lines within the Jurkat cell line can reveal the heterogeneity of tumour cells within the context of T-ALL.

Within the ten hallmarks of cancer, high proliferation is associated with aggressiveness and poor clinical outcomes. (Fouad and Aanei, 2017; Hanahan and Weinberg, 2011). Within cancer cells, driver mutations can allow for growth-factor dependent activation through dysregulated cell cycle processes. When this ability of cell cycle control is lost, tumours harbour the ability to bypass the tightly-regulated processes of the cell cycle and hyper-proliferate (Gutschner and Diederichs, 2012).

The uncontrolled proliferation of cancer cells reflects changes in the genetic and epigenetic landscape in the cancer cell (Gutschner and Diederichs, 2012). Different genetic and epigenetic alterations have been shown to give rise to different proliferative phenotypes for subtypes of T-ALL (Navarrete-Meneses and Pérez-Vera, 2017). Therefore, proliferation is a suitable phenotype to test for ideas of ITH within Jurkat clonal populations.

Cell proliferation assays are a simple technique to assess this phenotype amongst the derived Jurkat clones (Koyanagi et al., 2016). Flow cytometry-based analysis of

lymphocyte cell division uses the dye carboxyfluorescein succinimidyl ester (CFSE) (Azarsiz et al., 2018; Bocharov et al., 2013). This compound passively diffuses into cells and is not fluorescent until acetate groups are cleaved by esterases within the intracellular environment, resulting in fluorescent carboxyfluorescein succinimidyl ester (Azarsiz et al., 2018). CFSE stably labels intracellular proteins and the dye is equally partitioned between daughter cells during cell division. Therefore, each cell division results in each daughter cell having half the fluorescence intensity of the parental cell population (Azarsiz et al., 2018). This can be used in conjunction with real-time PCR (qPCR) expression analysis to test for phenotypic differences and transcriptional activity of T-ALL related genes.

The use of qPCR is predominantly done for the purpose of quantifying specific gene expression by amplification of cDNA (Kralik and Ricchi, 2017). Genes that are characteristic of the TAL1 CRC (*TAL1*, *GATA3*, *MYB* and *RUNX1*) can be quantified for expression within Jurkat clonal cell lines to further characterise the gene expression profiles of these cell lines.

The rationale of the study described here is that the generation of clonal populations of T-ALL Jurkat cells could be a way to understand ITH by identifying different clonal cell lines with different phenotypes that can be linked to genetic and/or epigenetic changes in each cell line. This chapter will describe the use of the CFSE assay to characterise the proliferation of the parental Jurkat T-ALL cell line as well as clonal cell lines derived from this cell line. This is linked to the analysis of gene expression to correlate proliferation with the expression of genes found within the TAL1 CRC.

2.2. Methods and Materials

2.2.1. Jurkat Cell Thawing and Culturing

Jurkat Clone E6-1 (ATCC TIB-152) was used to generate clonal populations. Jurkat cells were retrieved from cryogenic storage and immediately placed in a bead bath at 37°C to thaw. Complete media (Gibco RPMI 1640 GlutaMAX + 25mM HEPES media supplemented with 10% Foetal Bovine Serum (FBS) and 1% Penicillin-Streptomycin-Fungizone (PSF) (Gibco)) was pre-warmed to 37°C. Once cells were thawed, cells were aliquoted into 15mL Falcon tubes and 6mL of pre-warmed complete media was added to the tube within a sterile Biological Safety Class II cabinet. Cells were centrifuged at 300 x g for 5 minutes, supernatant was discarded, and cells were resuspended in 7mL of complete media and aliquoted into a T25 Cellbind Surface flask (Corning). Cells were then cultured in a humidified incubator at 37°C and 5% CO_2.

2.2.2. Jurkat Cell Clonal Population Generation

Thawed cells were passaged 2-3 times and clonal populations were isolated from the parental cell line (P0). Another parental population (P00) was also established as one passage different from P0. To do this, cells from the parental Jurkat cell line were resuspended at a concentration of $1x10^6$/mL in RPMI 1640 complete media and 10µL was removed and diluted into 990µL of complete media. This was used to prepare two serial dilutions of 10 and 1 cell/200µL in a volume sufficient for plating 200µL into each well of a 96-well round-bottom plate (typically one plate was divided equally between the two serial dilutions). Each 96-well plate was left to grow for up to 2 weeks to allow for growth of clonal populations. Although the 10 cells/200µL dilution increased the likelihood of generating a clonal cell line, analysis of cell lines generated from the 1 cell/200µL concentration were prioritised.

Cell populations were grown to confluency based on appearance within the field of view (80% of FOV) within the 96 round-bottom plate and then passaged into a larger volume 12-well plate (Corning). Once cells reached a cell concentration of at least $1x10^5$/mL (Total of $1x10^6$), cells were re-passaged into T25 flasks. Jurkat populations were then frozen down at a concentration of $1x10^6$/mL with 10% DMSO at -80°C for at least 24 hours, then placed within cryogenic storage.

2.2.3. Jurkat Clonal Populations

A single stock of vials for each parental cell line and each clonal cell line was created and stored within the cryogenic storage and all experiments were conducted by taking one vial from this stock. Jurkat cell lines brought up from the storage were considered passage 1 after thawing. This was done as Ben-David et al. (2018) found that freeze-thawing did not induce significant genetic variation. In this way, we generated 13 Jurkat cell lines. This included the two parental cell lines (P0 and P00) and clonal cell lines 1-11 (C1-C11). However, due to the lack of growth of C7, it was excluded from any further analysis in all experiments done within this project. Therefore, the functional cell lines used were P0, P00, C1, C2, C3, C4, C5, C6, C8, C9, C10 and C11.

2.2.4. Carboxyfluorescein Succinimidyl Ester (CFSE) assay

CFSE (eBioscience ThermoScientific) was reconstituted with 90µL of anhydrous Dimenthyl Sulfoxide (DMSO) (Sigma Aldrich) for a stock concentration of 10mM. A CFSE stock dilution to 10µM was done by aliquoting 50µL of 10mM CFSE into 49.05mL of Dulbecco's phosphate-buffered saline (PBS) (Gibco). This was further diluted to 2µM and 1µM CFSE by aliquoting 10mL or 5mL into a final stock of 50mL with PBS, respectively. This was stored at -20°C covered in foil to avoid fluorescent bleaching.

Jurkat cell lines were grown in complete RPMI 1640 media (10% FBS, 1% PSF) until passages 1, 5 and 9. Cells were then harvested and resuspended in 0.1% FBS and 1% PSF RPMI 1640 media at a concentration of $1x10^5$/mL for a 24 hours starvation.

The CFSE experiments were done using $5x10^4$ cells/well, with additional cells required from the parental P0 cell line for CFSE negative and 0-hour controls, and for cells to be stained with propidium iodide (PI) (Invitrogen). Cells were washed once in 1 x PBS by centrifugation at 300 x g for 5 minutes in a 15mL Falcon tube. To label cells with a saturating concentration of CFSE, cells were then resuspended at a concentration of $5x10^4$ cells/200µl in 1 x PBS and mixed with an equal volume of 1µM CFSE (final concentration of 0.5µM CFSE) and incubated at room temperature covered in foil for 10 minutes (as recommended by ThermoScientific). CFSE labelling was stopped by the addition of 4-5x volumes of 10% FBS, 1% PSF RPMI 1640 media (room temperature) and placed on ice for 5 minutes protected from the light. Cells were then centrifuged at 300 x g for 5 minutes, supernatant was removed, and cells were washed 3 more times. Cells were then resuspended in pre-warmed (37°C) 5% FBS, 1% PSF RPMI 1640 media at a concentration of $5x10^4$ cells/200µl and aliquoted at 200µL per well into the appropriate number of wells in a 96 well round-bottom plates. Positive (0-hour CFSE) controls were aliquoted at 200µL into 1.5mL centrifuge tubes. Cells were then incubated at 5% CO_2 at 37°C for 48 and 72 hours until flow cytometry analysis.

Negative (no CFSE) controls and 0-hour CFSE (positive) controls were run through the BD Biosciences Accuri C6 Plus flow cytometer to establish negative and positive gates to be used for CFSE analysis. The BD Accuri C6 Plus flow cytometer contains a 488 nm (blue) and 640nm (red) laser with 4 detectors. Filters FL1-FL4 detect ranges of wavelengths of 533/30, 585/40, 670LP and 675/25, respectively as stated by BD Biosciences

(https://www.bdbiosciences.com/documents/BD_Accuri_C6Flow_Cyto_Instrument_Manual.pdf). For CFSE detection, FL1 was used as CFSE has an emission spectrum of 521nm (Thermo Fisher Scientific) and the PI stain (Invitrogen) was detected using the FL3 detector for an emission of 493/636nm.

2.2.5. Propidium Iodide (PI) Staining

Jurkat cells that were grown over 48 or 72 hours within 5% FBS and 1% PSF of complete RPMI 1640 media were harvested. Cells were then washed twice by adding 2mL of PBS, centrifuging at 300 x g for 5 minutes and decanting the supernatant. Cells were then resuspended in 100µL of staining buffer (5% FBS in PBS) and mixed gently with 5µL of PI (250µg/mL) (Invitrogen) and incubated for one minute and avoiding light. Fluorescence of PI was then determined by flow cytometry as stated above.

2.2.6. CFSE Analysis

Jurkat cell lines were run through the BD Accuri C6 Plus Flow Cytometer and gated based on no CFSE, 0-hour CFSE controls and PI cell death stain controls. A limit of 10,000 events was placed on the analysis per cell line (if possible) to avoid over saturation of fluorescence due to a large cell count. FCS files from the BD Accuri C6 plus program were extracted and used to visualise results in the FlowJo Program (Version 10).

FlowJo is a commonly used program for analysing and visualising Flow Cytometry data in an easy-to-use interface (Lugli et al., 2010; Quah et al., 2007). Raw overlay histograms were generated using FlowJo to display positive controls, negative controls and CFSE samples at 48 hours and 72 hours. The FlowJo proliferation module identifies patterns expressed by cells loaded with fluorescent dyes, such as CFSE. Using optimised algorithms, it models cell division based on gates created around the original 0-hour control population and conducts a peak analysis.

The FlowJo proliferation module outputs various values including: division index: average number of cell divisions that all cells have undergone, proliferation index: the average number of cell divisions that the 'responding' cells have undergone, expansion index: the fold-expansion of the overall population and replication index: fold-expansion of the 'responding' cells. The term 'responding cells' refers to the cells in the population that have undergone cell division at least once and is detected by FlowJo. FlowJo have recommended that the proliferation index (PI) is the most appropriate value to analyse as it is the closest to analysing proliferation processes biologically. Jurkat cell lines at specific passages were run through the proliferation module and gated based on the starting population and adjusted for optimal model fit of peaks through changing the option of the number of peaks within the module interface. Values were extracted and placed into .csv files and analysed per cell line and per passage.

CFSE data was visualised using the R package and online platform, Plotly (https://plot.ly) to generate boxplots of PI values amongst clonal and parental cell lines. Data was further analysed using Socscistatistics Mann-Whitney U calculator (https://www.socscistatistics.com/tests/mannwhitney/). The Mann-Whitney U Test was used as a non-parametric test to see if sample values originate from the same distribution and compare between two independent samples (parental and clonal cell line populations) (Nahm, 2016).

The data was also analysed by the Astatsa Kruskal-Wallis rank sum calculator (https://astatsa.com/KruskalWallisTest/) to identify statistical differences amongst multiple samples and further adjusted for multiple comparisons using a Dunn's test (Nahm, 2016). The Dunn's test was utilised as a Post-Hoc test after Kruskal-Wallis testing and recommended for its ability to identify precise statistical differences and

account for multiple sample comparisons (Lee and Lee, 2018). It is further recommended to use a Holm Family-Wise Error Rate (FWER) or Benjamini-Hochberg false discovery rate (FDR) p-value adjustments for rigorous testing (Lee and Lee, 2018)

For box plot analysis of mild and extreme outliers, Plotly was used to retrieve values of the lower quartile (Q1) and the upper quartile (Q3). These values were used to calculate the inner and outer, upper and inner fences to establish thresholds to determine groups of proliferative ability of clonal cell lines at specific passages. The following calculations were used:

Interquartile range (IQ): Q3 – Q1

Lower inner fence: Q1 – 1.5 x IQ

Upper inner fence: Q3 + 1.5 x IQ

Lower outer fence: Q1 – 3 x IQ

Upper outer fence: Q3 + 3 x IQ

The Socscistatistics Friedman test calculator (https://www.socscistatistics.com/tests/friedman/default.aspx) was also used along with the Socscistatistics Wilcoxon-Signed Rank test calculator (https://www.socscistatistics.com/tests/signedranks/default.aspx). A Friedman Test was conducted due to its ability to assess non-parametric data and compare differences between multiple repeated measures (such as passages). This utilises a procedure called the 'method of ranks', similar to other non-parametric tests, and does not assume normality or homogeneity of variance amongst samples (Eisinga et al., 2017). A pairwise comparison can then be conducted if differences are found within the Friedman test using the Wilcoxon-Signed Rank as a Post-Hoc test (Kim, 2014).

2.2.7. RNA isolation

As described in sections 2.2.2 and 2.2.4, Jurkat cell lines (P0, P00, C1-C6 and C8-C11) were generated and tested for rate of proliferation at passages 1, 5 and 9. This was done after a 0.1% FBS and 1% PSF starvation in RPMI 1640 media for 24 hours and cells were then grown in 5% FBS 1% PSF RPMI 1640 media for 48 and 72 hours (Methods: 2.2.1-2.2.4). In order to test the gene expression profiles of these Jurkat cell lines under the same treatment, Jurkat cell lines were grown under the same conditions for 72 hours and RNA was then harvested for gene expression analysis.

RNA isolation was done using the Bioline RNA ISOLATE II Mini Kit (https://www.bioline.com/us/downloads/dl/file/id/885/isolate_ii_rna_mini_kit_product_manual.pdf). Purification was done according to the manufacturers' instructions, with the exception that prior to the elution of RNA, an extra two-minute centrifugation at 11,000 x g was done with no added wash buffer. RNA was then eluted with 60µL of double-distilled H_2O (ddH_2O) and centrifuged at 11,000 x g for one minute. Eluted RNA was then measured for concentration using the Qubit 2.0 Fluorometer using the Qubit 2.0 Fluorometer BR assay kit (ThermoFisher) as per the manufacturers' instructions.

2.2.8. cDNA Preparation

For cDNA synthesis, 50ng of RNA was used in conjunction with a 5x TransAmp Buffer (Bioline) at a final concentration of 1x, and 1µL of reverse transcriptase (Bioline) in a total of 20µL. cDNA was synthesized using a thermocycler at: 25°C for 10 minutes, 42°C for 15 minutes, 85°C for 5 minutes and hold at 4°C. The synthesised cDNA reactions were then diluted at a 1:1 ratio to have a final volume of 40µL to be used for qPCR gene expression analysis.

2.2.9. Primer Design

Primers were designed based on the following method. Initially, the NCBI gene search engine (https://www.ncbi.nlm.nih.gov/gene/) was used to identify gene mRNA transcripts. Exon sequences that spanned at least one intron were copied into the Primer3 program (http://bioinfo.ut.ee/primer3-0.4.0/) with the following parameters: Primer Length: 19-21, T_m: 59-61°C with a max T_m Difference of 1°C. Primer sequences outputted from Primer3 were copied into the UCSC database through the 'In-Silico PCR' function (https://genome.ucsc.edu/cgi-bin/hgPcr). If targeted amplicons were identified in the UCSC database and showed mapping of primers to amplify exon sequences that span an intron, the primer was used. The following primer sets were made for the genes, *TAL1*, *GATA3*, *RUNX1*, *MYB* and *NFYB* (Reference Gene).

| Table 2.1. Sequences for Forward and Reverse Primers of the genes, *TAL1*, *GATA3*, *MYB*, *RUNX1* and *NFYB*.

Gene[1]	Forward Primer 5'-3'	Reverse Primer 5'-3'
TAL1	CCCCCTATGAGATGGAGATT	AAAGGCCCCGTTCACATT
GATA3	CTCTCTGCTCTTCGCTACCC	GCGACGACTCTGCAATTCT
MYB	TGGACCAAAGAAGAAGATCAGA	TCTCCCCTTTAAGTGCTTGG
RUNX1	ACTCGGCTGAGCTGAGAAATG	GACTTGCGGTGGGTTTGTG
NFYB	AGATTGCAAAAGATGCCAAAG	CGTTTCTCTTGATGGCACCT

[1]Primer amplicon sequence information in Appendix – Supp. Table 7.7 and section 7.1.2.a.

Primer efficiencies were tested with Jurkat cDNA at a starting concentration of 5ng with sequential 1:1 dilutions to 0.3125ng (5, 2.5, 1.25, 0.625 and 0.3125ng). Primer efficiencies were tested in triplicate at these concentrations and cycle threshold values (CT values) were extracted from qPCR analysis (Fig 2.9, y-axis). A standard curve of best fit was placed on all data points and the equation of the slope was derived and used to calculate for efficiency using the 'qPCR Primer Efficiency Calculator' by

ThermoFisher Scientific (https://www.thermofisher.com/au/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/qpcr-efficiency-calculator.html). Primer efficiencies of 90-110% are ideal for conducting qPCR analysis (Rogers-Broadway and Karteris, 2015).

2.2.10. Quantitative Analysis of Gene Expression (qPCR)

The qPCR gene expression reactions were done using the SensiFAST SYBR No-ROX Kit (Bioline) that provides a 2x Master Mix containing SYBR Green I dye, dNTPs and qPCR stabilisers and enhancers (BioLine: https://www.bioline.com/downloads/dl/file/id/2754/sensifast_sybr_no_rox_kit_manual.pdf). qPCR reactions were conducted in a final volume of 10µL that included 2x SYBR No-ROX (Bioline) at a final concentration of 1x, 2µL of diluted cDNA, and forward and reverse primers at a final concentration of 0.2µM each.

qPCR reactions were carried out using an Applied Biosystems 7500 Fast Real-Time PCR System with the following cycling conditions: Initial denaturation at 95°C for 20 seconds, cycling stage of 95°C for three seconds and 60°C for 30 seconds for 40 cycles. This was followed by a melt curve analysis with the following cycling conditions: 95°C for 15 seconds, 60°C for 60 seconds, 95°C for 15 seconds and 60°C for 15 seconds. After analysis was done, CT values were extracted and used for gene expression analysis.

2.2.11. qPCR Gene Expression Analysis

Gene expression was analysed using the 2^{-ddCT} method (Rao et al., 2013) whereby CT values of target genes were normalised to a reference gene (dCT) and then relative to a control sample (ddCT). Fold difference was then calculated by using the 2^{-ddCT}

formula. This formula assumes amplification of PCR products with 100% efficiency (Rao et al., 2013) which was confirmed prior to the analysis of gene expression.

qPCR fold change data was visualised using the R package, 'Pheatmap' (https://www.rdocumentation.org/packages/pheatmap/versions/1.0.12) to generate the Euclidean Distance Hierarchical Clustering heatmap of all Jurkat cell lines at all passages (Command line in Appendix – section 7.1.2.b). Data was further analysed using Socscistatistics Friedman Test calculator (https://www.socscistatistics.com/tests/friedman/default.aspx) and the Wilcoxon-Signed Rank Test (https://www.socscistatistics.com/tests/signedranks/default.aspx). The data was also analysed by the Astatsa Kruskal-Wallis rank sum calculator (https://astatsa.com/KruskalWallisTest/).

A permutation test of gene sets is widely used in genomic research as it provides statistical inference with assumptions not being based on strict distributions of data (Bůžková et al., 2011). It conducts analysis through the comparison of observed values to the distribution of data from groups of permutations that would not affect the distribution if no difference between groups was observed (Bůžková et al., 2011). Despite permutation testing being widely used, it is computationally-demanding and the speed of the analysis can be compromised (Chang and Tian, 2016). Therefore, the package 'GSALightning' (https://github.com/billyhw/GSALightning) within the R programming language was used for this analysis. The GSA-Lightning package computes student-T statistics for each gene set and provides gene set statistics of significance (Chang and Tian, 2016). Command line in Appendix – section 7.1.2.c.

2.3. Results

2.3.1. Optimisation of CFSE assay using the Jurkat Clonal Populations

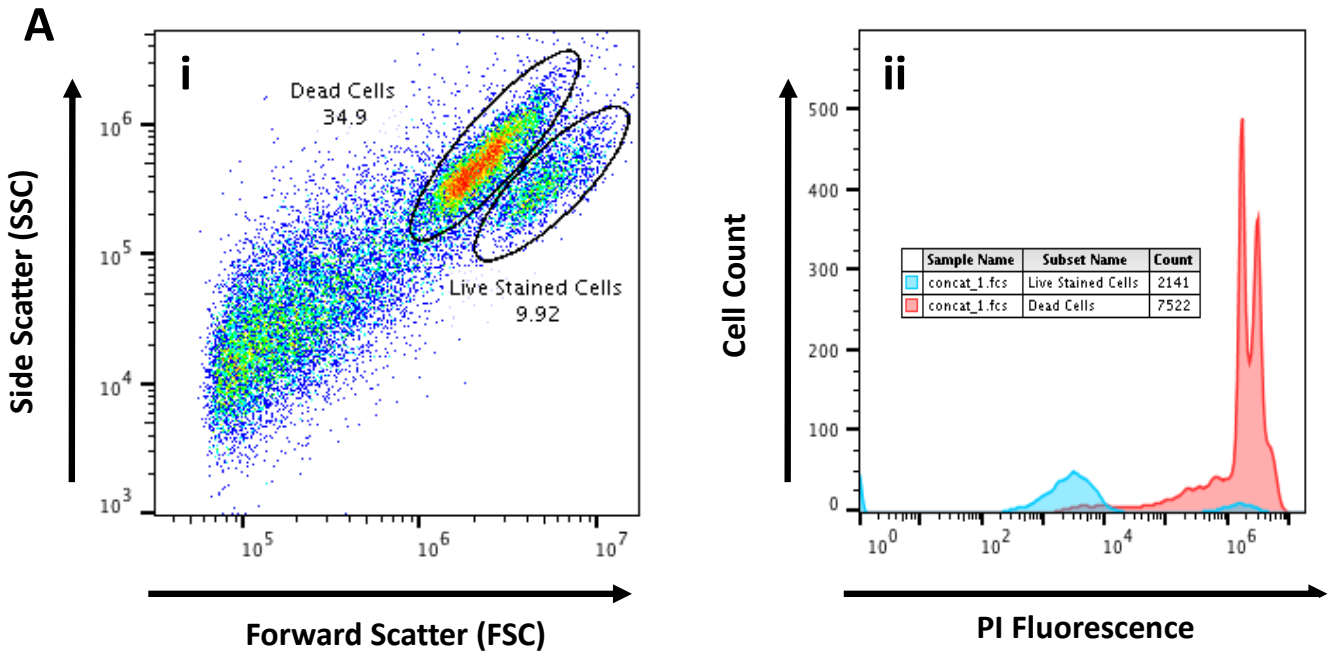
Typically, CFSE staining of lymphocytes uses cell concentrations ranging from $0.5x10^6$ – $1x10^8$/mL, with the concentration of CFSE ranging from a final of 0.5µM to 5µM,

dependent on the brand of CFSE used (Quah et al., 2007). It must be thoroughly resuspended in fresh media to buffer the toxic effects of the CFSE (Quah and Parish, 2010). Therefore, optimising the ratio of cell concentration to CFSE concentration is essential to maximise cell labelling while avoiding the toxic effects of CFSE (Quah et al., 2007). Furthermore, the fluorescence intensity of CFSE is not stable within the first 24-48 hours of an assay and labelled cells should be cultured for a minimum of 18 hours before analysis, although changes in cell proliferation are not typically seen until after 48 hours of treatment (Quah et al., 2007). In the experiments described here, cells were left to grow for up to 72 hours to ensure valid detection of fluorescence.

Initially, varying cell concentrations and CFSE concentrations were tested. The parental Jurkat cell line P0 was initially tested at a concentration of $2x10^4$ cells/well. A lower concentration of cells was grown relative to that cited in the literature due to our use of 96-well round-bottom plates that have a cellular capacity of $2\text{-}10x10^4$ cells/well, and to ensure equal labelling of CFSE dye to cells as stated by ThermoFisher Scientific (https://www.thermofisher.com/au/en/home/references/gibco-cell-culture-basics/cell-culture-protocols/cell-culture-useful-numbers.html).

A separate control group of cells were permeabilised and stained with PI stain to detect a PI-positive population that represents dead cells (Fig 2.1A, i and ii). This allowed for the gating of the live cell population within the tested populations in conjunction with FSC and SSC parameters for size and granularity (Fig 2.1, i). Prior to labelling testing populations with CFSE, cells were starved in 0.1% FBS for 24 hours. This was required as Jurkat cells maintained in 10% FBS proliferated rapidly, diluting the CFSE out of the cell population within the first 48 hours (Appendix – Supp. Fig 7.1). Therefore, cells were labelled at a final concentration of 0.5µM or 1µM of CFSE after starvation in 0.1% FBS and then cultured for up to 72 hrs in 0.5, 1, 2, 5 and 10% FBS (data not shown). It was found that at 5% and 10% FBS, a live cell population was

established relative to the 0.5-2% FBS which only displayed the dead cell population based on PI staining (data not shown).

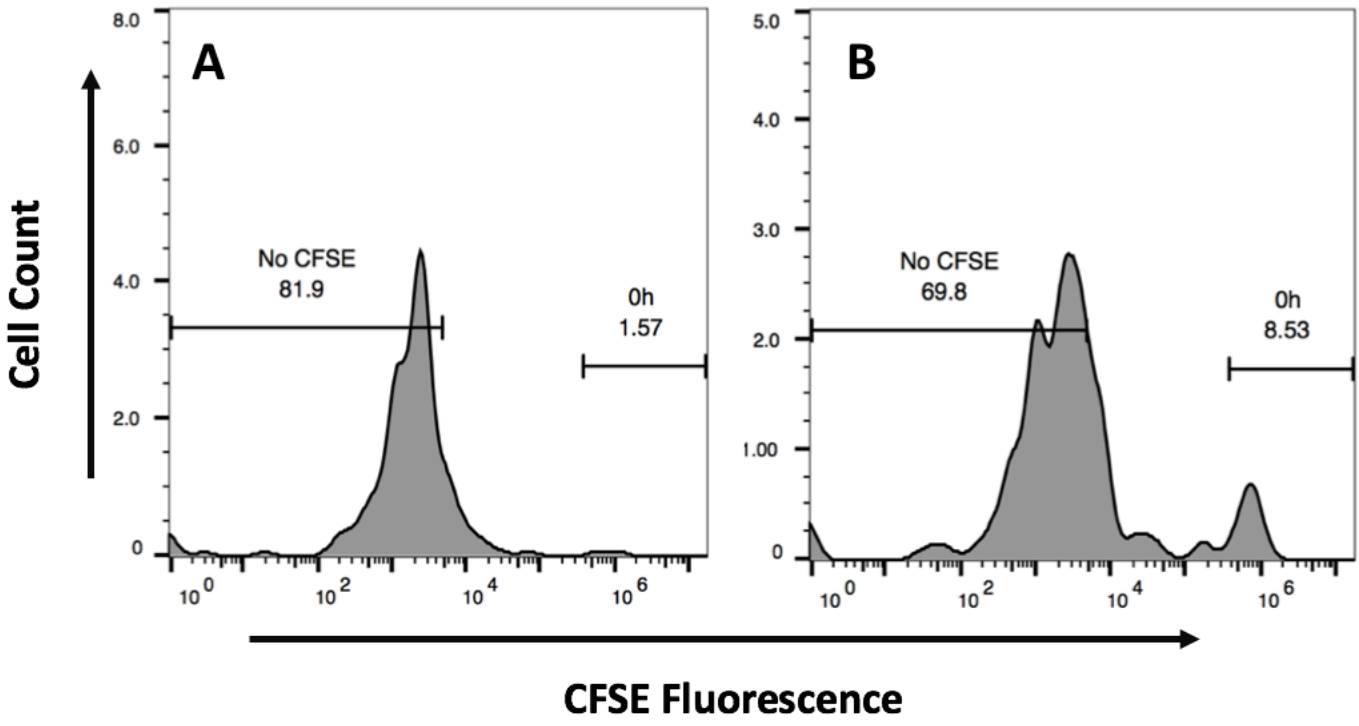


| Figure 2.1. Identification of Dead Cells using Propidium Iodide.

(**A)** Propidium Iodide (PI) stained permeabilised Jurkat cells (P0, passage 1) gated within the dot plot (**i**) and displayed in the histogram plot (**ii** - red) as validation of the accurate gating of the dead cell population. Dot plot consists of the x and y axis displaying FSC and SSC, respectively and displaying dead cell population of 34.9% and live stained cell population of 9.92% due to permeabilisation of Jurkat clonal cell lines. Histogram overlay plot displaying the dead cell gated population for fluorescence intensity of PI along the x-axis and cell count along the y-axis. Live stained and dead cells have a cell count of 2141 and 7522, respectively.

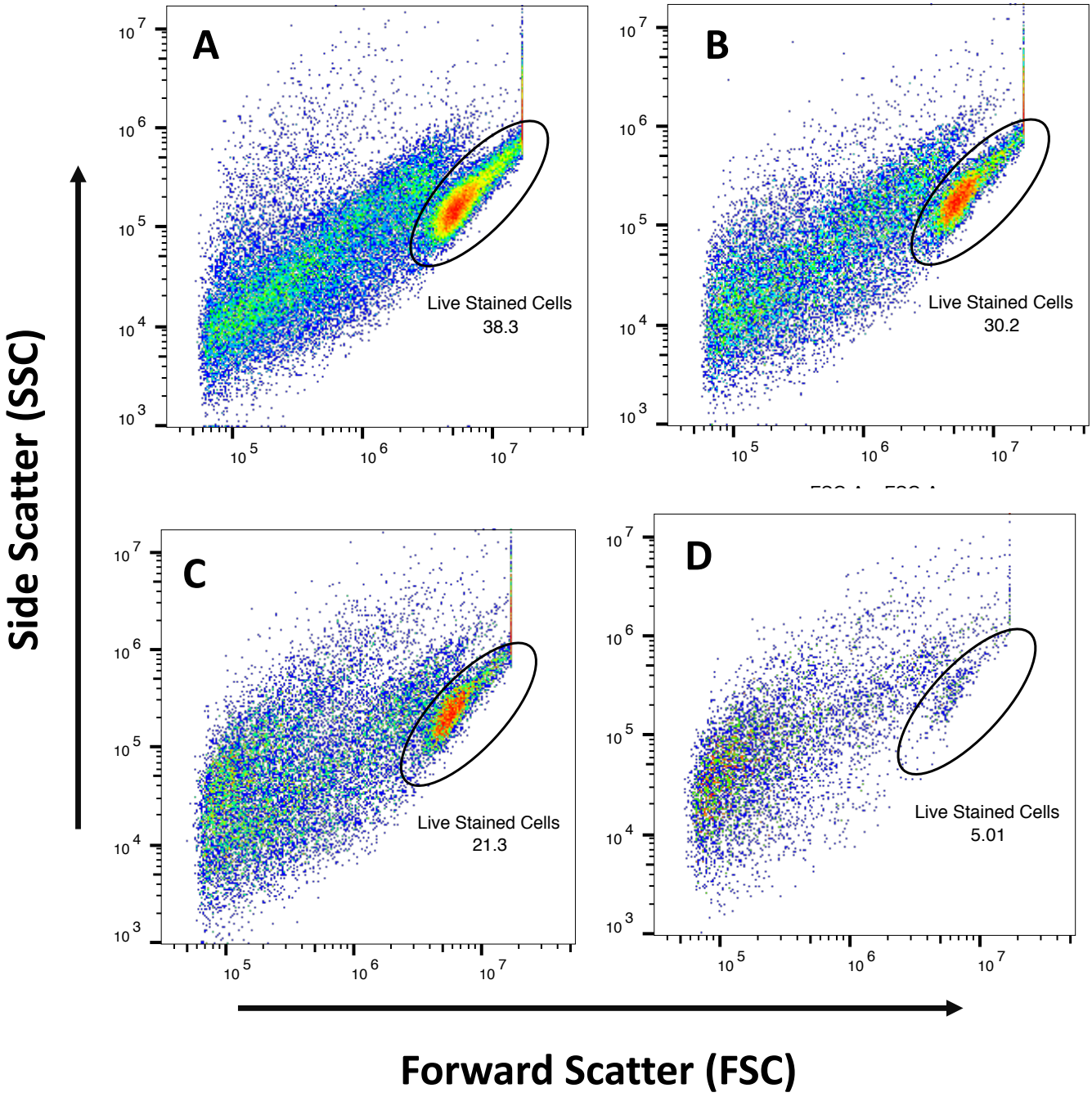
It was found that at 5% and 10% FBS, a live cell population was established relative to the 0.5-2% FBS which only displayed the dead cell population based on PI staining (data not shown and Fig 2.1). Therefore, further analysis of 5% and 10% FBS conditions were conducted for final 0.5µM and 1µM CFSE concentrations. Cell populations gated within 0.5µM and 1µM conditions varied in cell percentage at 5% FBS (19.6% vs 13.9% of the total population), similar to those seen using 10% FBS for duplicates tested (data

not shown). This suggests that CFSE a final concentration of 1µM becomes toxic to cells at concentrations of $2x10^4$/well, despite FBS concentrations of 5% and 10% within each cell population.



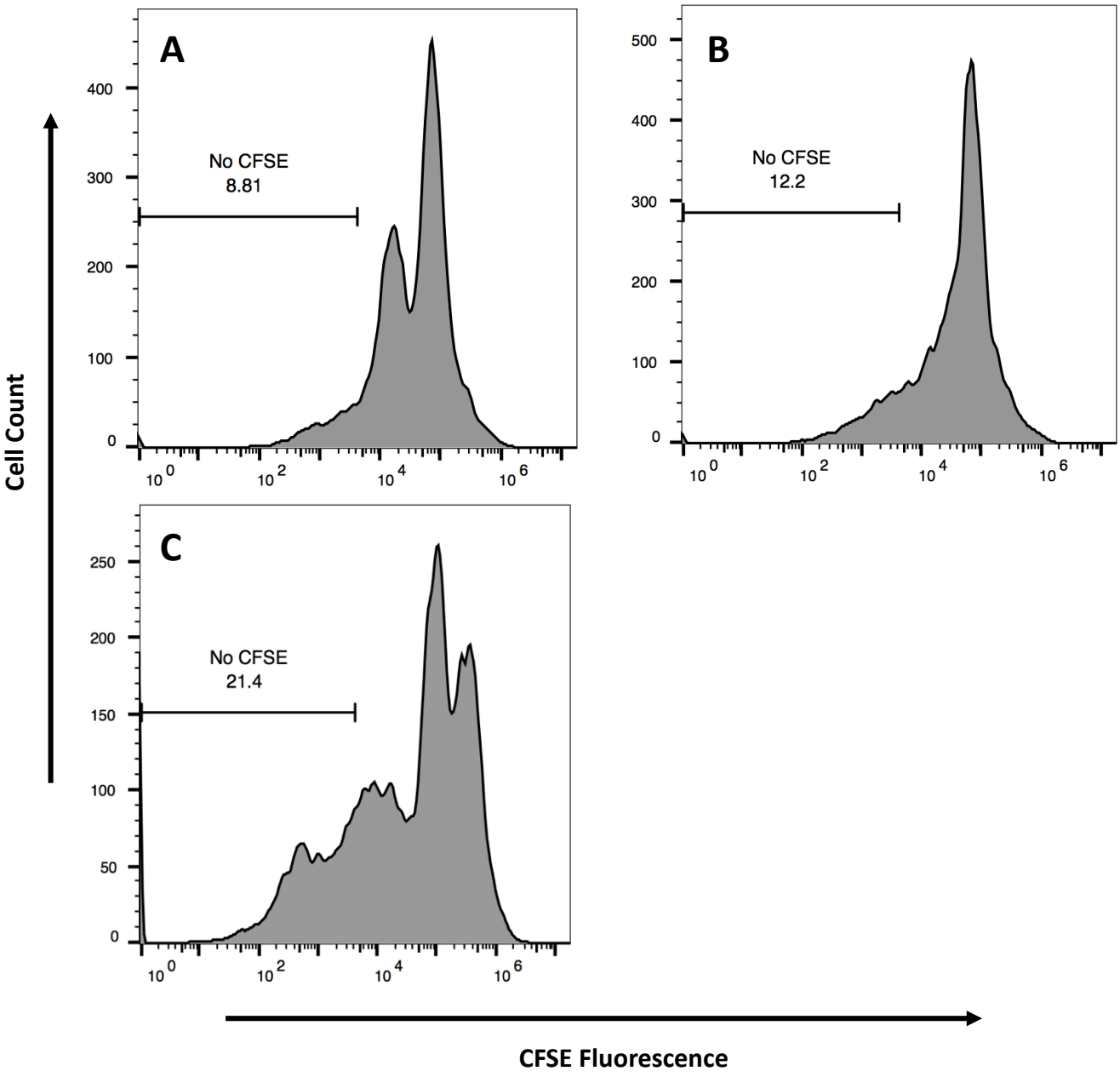
| Figure 2.2. Flow cytometry histogram plots of CFSE fluorescence at 0 and 72 hours post-labelling in the presence of 5% FBS.

(A-B). Negative and positive controls are gated from the overall population of cells to show the non-stained cells in the population (negative control) and CFSE-stained cells (positive control) (respectively). **A.** Negative control showing fluorescence from cells not labelled with CFSE and cultured for 72 hours post-starvation and positioning of no CFSE gate (81.9% of cells exhibit no CFSE). **B.** Fluorescence of positive CFSE loaded cells and cultured for 0 hours after staining and positioning of CFSE stained gate (8.53%). Histogram plot of CFSE fluorescence immediately after labelling cells with a final concentration of 0.5µM CFSE and 0-hour CFSE-positive gate.



| Figure 2.3. Flow-cytometry dot plots demonstrating the forward scatter (FSC) (x-axis) and side scatter (SSC) (y-axis) of Jurkat Cells to distinguish cell populations.

Cells were grown over 72-hours with a final CFSE concentration of 0.5µM and tested with varying cell concentrations: **A.** $2x10^5$/well (38.3% live cells); **B.** $1x10^5$/well (30.2% live cells); **C.** $5x10^4$/well (21.3% live cells) and **D.** $2x10^4$/well (5.01% live cells). Gated cell populations (circle) display percentage of cells that are predicted to be live within the entire cell population based on size and granularity (FSC and SSC) and PI staining of dead cells (Fig 2.1).



| Figure 2.4. Flow cytometry histogram plots of CFSE stained Jurkat Clone 2 at cell concentrations of $2x10^5$, $1x10^5$ and $5x10^4$/well.

(**A-C**, respectively) after 0.1% FBS starvation for 24 hours and grown in 5% FBS for 72 hours and stained with a final concentration of 0.5µM CFSE. Gates were established based on the negative control (no CFSE stained sample) and display the percentage of cells within the total population that fall within the fluorescence of the negative control (8.81, 12.2 and 21.4% respectively). Cell count is demonstrated on the x-axis and CFSE fluorescence intensity is plotted logarithmically along the y-axis.

Flow cytometry histograms were graphed and gated based on CFSE unlabelled negative controls from the live cell gated populations and 0-hour CFSE-labelled positive controls (Fig 2.2,A and B, No CFSE and 0-hour, respectively). In a CFSE proliferation experiment, CFSE-positive cell populations should be visible as multiple peaks that represent differences in the proliferation of cells over the course of an experiment (Quah et al., 2007). At 5% FBS, a clearer resolution for multiple peaks were detected relative to the 10% FBS condition for $2x10^4$/well (22% vs 9.58% - data not shown). However, despite higher cell numbers being available for analysis when using a final concentration of 0.5μM CFSE and 5% FBS, approximately 80% of cells were either not stained or fell within the dead cell population (data not shown). This led to further testing using different cell concentrations with 5% FBS and a final concentration of 0.5μM CFSE.

Having observed that the outcome of a CFSE labelling experiment is dependent upon the concentration of FBS in the culture media, we next verified the optimal cell concentration to be used for this assay. Based on PI staining (Fig 2.1), live cell populations were gated with the percentage of live cells differing between cell concentrations: 38.3, 30.2, 21.3 and 5.01% of the total cell population at concentrations of $2x10^4$, $1x10^5$, $5x10^4$ and $2x10^5$ cells/well, respectively (Fig 2.3 – percentage in gated circle). At a cell concentration of $2x10^4$/well, only 5% of cells were alive and therefore omitted from further analysis (Fig 2.3D – 5.01%), whereas all concentrations above this ($2x10^5$/well, $1x10^5$/well and $5x10^4$/well) displayed a range of 21-38% of cells being stained and alive (Fig 2.3A-C – 38.3%, 30.2% and 21.3%, respectively). Therefore, a histogram peak analysis of proliferation was done for cells grown at $5x10^4$/well, $1x10^5$/well and $2x10^5$/well (Fig 2.4).

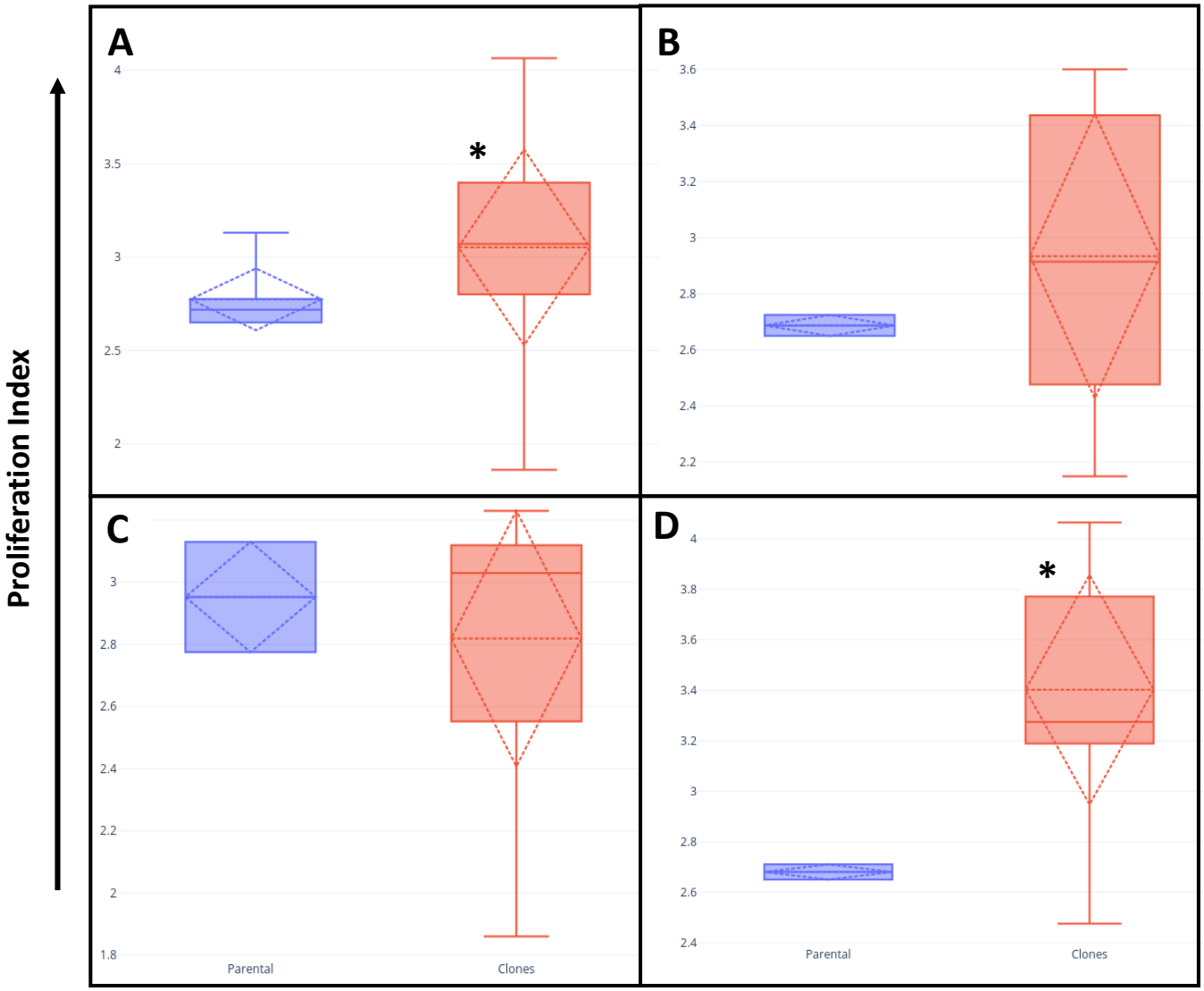
Using histogram peak analysis, cell concentrations of $2x10^5$ and $1x10^5$/well displayed a minimal number of CFSE peaks, and an absence of different proliferating cell

populations (Fig 2.4A,B). In contrast, a cell concentration of $5x10^4$/well displayed multiple CFSE peaks that represented the differences in the proliferation of cell populations over 72 hours (Fig 2.4C). Therefore, it was concluded that prior to labelling with CFSE, cells would be starved in 0.1% FBS for 24 hours, labelled with a final concentration of 0.5µM CFSE, and plated at a concentration of $5x10^4$ cells/well in 5% FBS, and assayed at for 0, 48 or 72 hours.

2.3.2. Analysis of proliferation index between Jurkat parental cell lines and the clonal cell lines

The CFSE assay was used to test for differences in proliferation between clonal populations and in comparison, to the parental Jurkat populations (P0 and P00). As earlier work has shown cell passage number is an important determinant of genetic and phenotypic variability within cultured cell lines (Ben-David et al., 2018), and as the rationale for experiments described here is the correlation between phenotypic variability with genetic and/or epigenetic variability, all cell lines were grown continuously and analysed at passage numbers 1, 5 and 9.

First, a comparison of proliferation between parental and clonal cell lines (C1-C6 and C8-C11) was conducted and plotted as a box plot using proliferation index (PI) values across all passages combined (Fig. 2.5A), as well at passages 1, 5 and 9 individually (Fig 2.5, B-D, respectively). No significant difference was found between parental and clonal cell lines for PI at passages 1 and 5 (Fig. 2.5, B and C, $p > 0.05$, Mann-Whitney 2-tailed U-test, see also Table 2.2), although significant differences in PI between parental and clonal cell lines were found when all data for all passages was combined, and at passage 9 only (Fig. 2.5, A and D respectively, $p < 0.05$, Mann-Whitney 2-tailed U-test, see also Table 2.2). These results showed that parental and clonal cell lines are similar in terms of their proliferative ability within the first 5 passages but differ in their proliferative ability at passage 9, which also accounted for the significant difference in PI when all passage number data was analysed together.



| Figure 2.5. Proliferation indices for Jurkat parental and clonal cell lines.

A. Proliferation indices for parental Vs clonal cell lines for all passages combined. **B-D**. Proliferation indices for parental Vs clonal cell lines at passage numbers 1, 5 and 9, respectively. Parental clone lines (P0 and P00) are represented in blue boxplots and the derived clones are displayed as red box plots. Passage 1 and 5 have a significance of $p > 0.05$ and passage 9 and all passages have a significance of $p < 0.05$ (*) as determined by a two-tailed Mann-Whitney U test. (Parental cell line data, n=12; clonal cell line data n=20).

2.3.3. Analysis of the proliferation index between different Jurkat clonal cell lines

Next, differences in proliferation between individual clonal cell lines relative to the parental cell lines were investigated by plotting duplicate PI values from passages 1, 5 and 9 (Fig 2.6). In comparison to the parental cell lines P0 and P00, and in particular when compared to P0, a large variation in PI between clones was seen across passages, such as C1, C5, C6, C9, C10 and C11 (Fig 2.6). This contrasts with clonal populations C3, C4 and C8 which showed a smaller variation of PI despite a higher median PI relative to the parental cell lines (P0 and P00) (Fig 2.6). Thus, the PI of each cell line per passage was tested using a Kruskal-Wallis rank test.

Statistically significant differences were seen when analysing all cell lines at all passages using the Kruskal-Wallis test ($p < 0.05$). However when p-value readjustments were conducted using the Holm FWER and the Benjamini-Hochberg FDR method, no statistical differences were seen between cell lines at all passages (Appendix – Supp. Table 7.1-7.6).

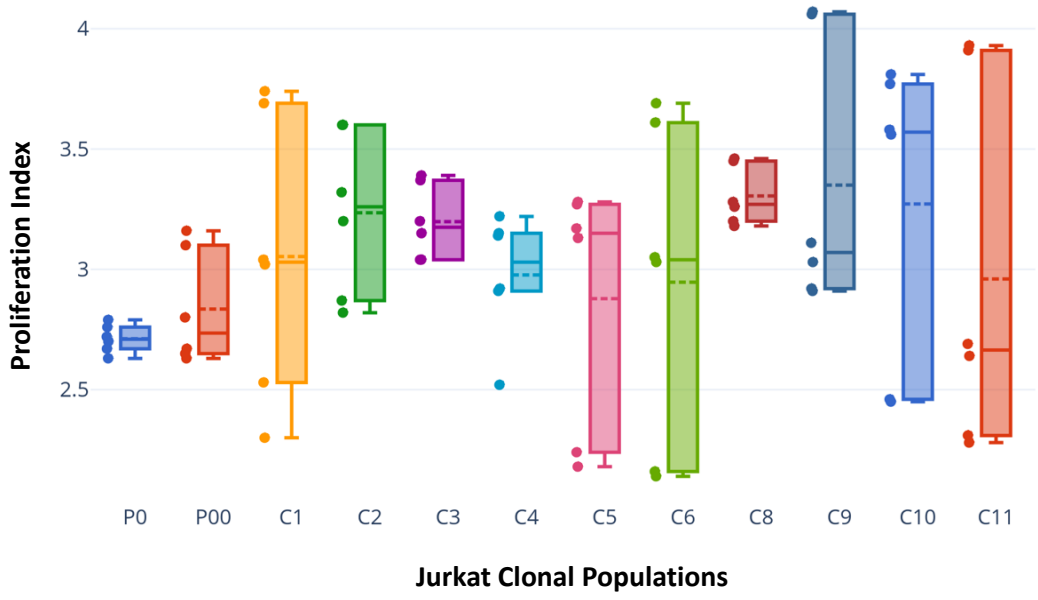
2.3.4. Subgrouping of Jurkat Clones based on a parental Jurkat cell line (P0) outlier analysis

The analysis of parental and clonal cell lines showed statistically different distributions of PI, but Post-Hoc testing was unable to identify differences between individual cell lines. This led us to consider a different approach to understand differences in PI between individual cell lines. Using box plot outlier analysis (Kwak and Kim, 2017), all cell lines at all passages were graphed and grouped based on outlier values within upper and lower inner and outer fences to determine subgroupings of proliferative ability. Upper and lower inner and outer fences were calculated based on the data distribution of the parental population (P0) (Methods: 2.2.6) as all Jurkat clonal populations are derived from this original population. This

displayed how clonal populations isolated from this original population differ in PI. Values between inner fences and outer fences are considered as mild outliers and

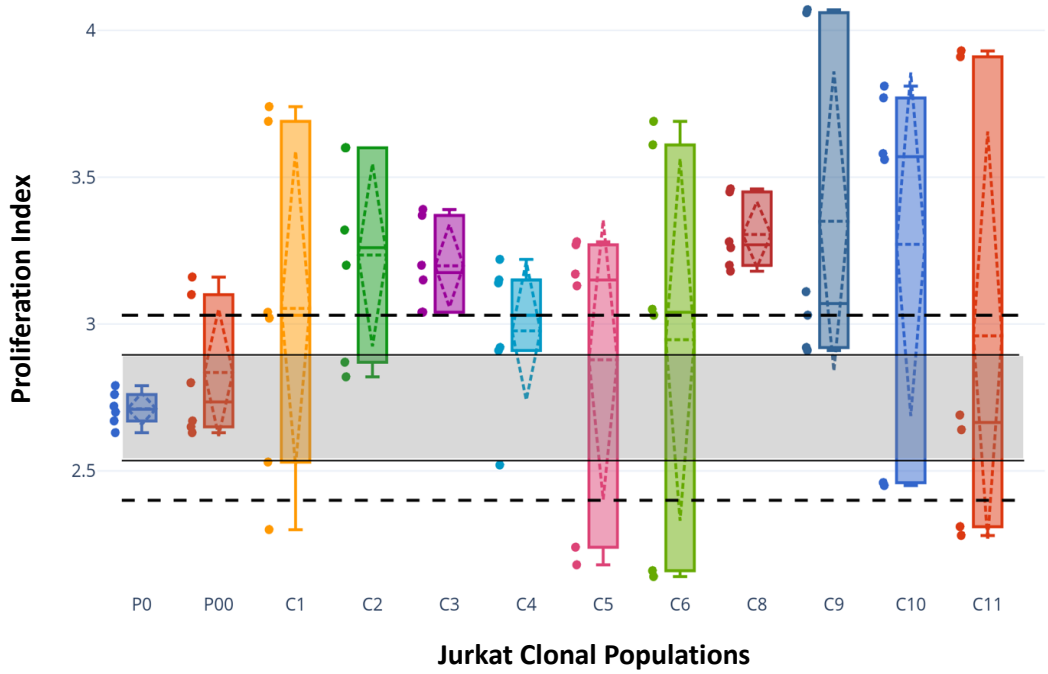
| Table 2.2. Mann-Whitney U test statistical test results of proliferation index (PI) of parental Jurkat cell lines (P0 and P00) and Jurkat clones (C1-C6, C8-C11).

Sample	p-value	Significance (p < 0.05)
All Passages	0.02	Yes
Passage 1	0.27	No
Passage 5	0.86	No
Passage 9	0.03	Yes



| Figure 2.6. Boxplot display of proliferative index (PI) duplicate values per clone (n=6) at passages 1, 5 and 9.

Data points are displayed on the left of the respective boxplot. Dashed lines and solid lines indicate the mean and median of data distributed per boxplot, respectively. There is a significant difference in the PIs across all cell lines (Kruskal-Wallis test ($p < 0.05$), see also Appendix – Supp. Table 7.1-7.6). However, using the Dunn Post-Hoc test with p-value adjustments by Holm Family-Wise Error Rate (FWER) method, statistical differences were not seen between individual cell lines.



| Figure 2.7. Outlier analysis of proliferative index for parental and clonal cell lines.

Shown are coloured boxplots of individual cell lines, in each case representing duplicate PI values for each of passage 1, 5 and 9 (n=6). The horizonal black dashed lines represent the lower and upper outer fences calculated from P0 PI distribution for all cell lines at all passages. The upper and lower boundaries of the grey shaded area show the upper and lower inner fences. The grey shaded area is the region within which PI data values are not considered an outlier based on the distribution of data from the parental Jurkat cell line (P0) as the clonal cell lines are derived from this original population. Data points are displayed on the left of the respective boxplot, with a display of the median and mean of PI as indicated by the solid and dashed lines, respectively. Standard deviation of the data per passage is indicated by the peaks of the dashed lines (±).

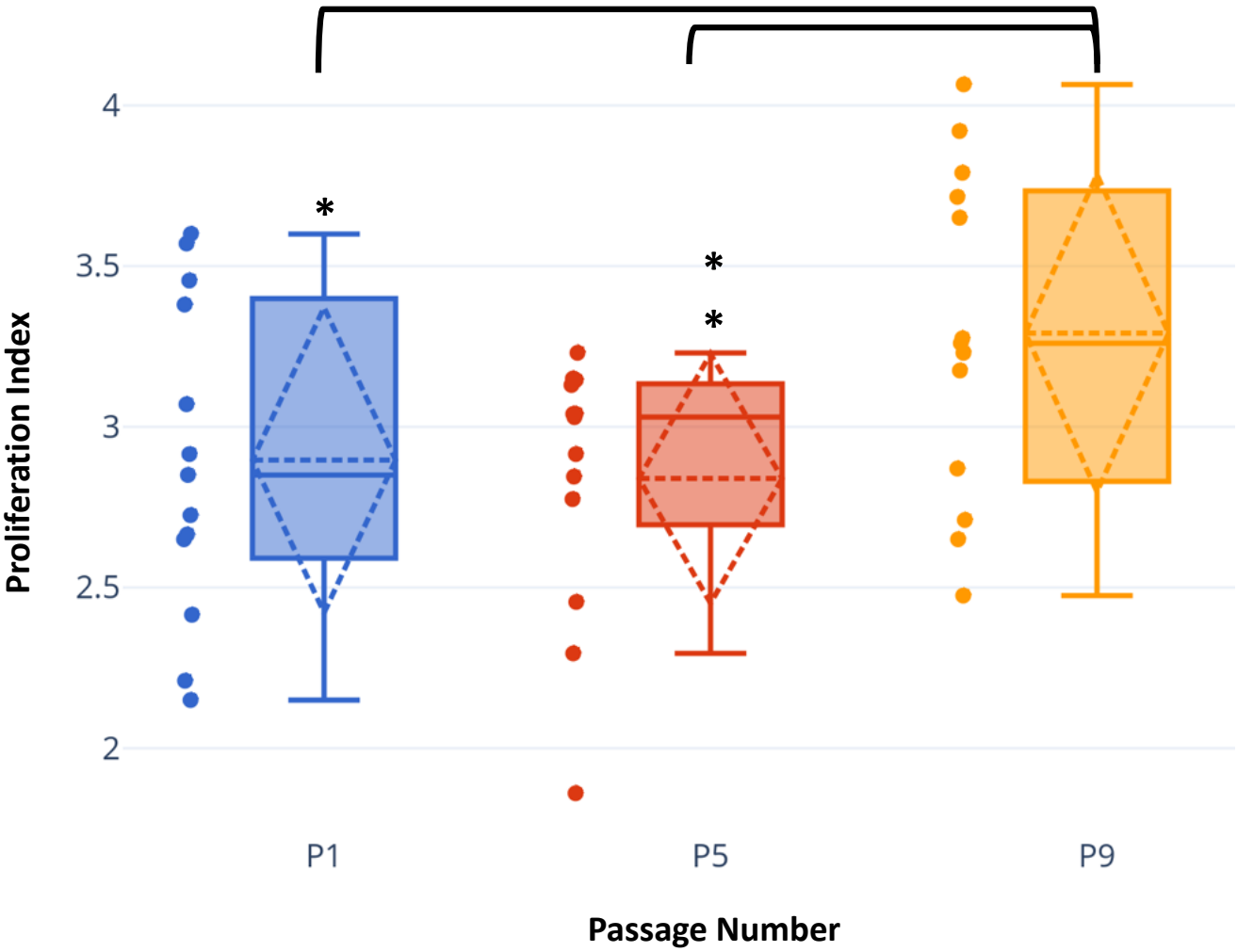
grouped as moderate high or low proliferation (Fig 2.7, values that fall between the upper or lower black dashed lines and the corresponding upper or lower boundary of the grey shaded region). Cell lines at all passages that fell outside of outer fences (Fig 2.7, values that above or below upper or lower black dashed lines, respectively) were considered as extreme outliers and grouped as extreme high or low proliferation. The results of this outlier analysis showed that the majority of clones were placed within

| Table 2.3. List of Jurkat cell lines categorised by their proliferative ability from the boxplot outlier analysis (Fig 2.7) (see also Methods: 2.2.6 for calculation method).

Category	Cell Line	Passage
Extreme High	C2	1
	C3	
	C8	
	C10	
	P00	5
	C1	
	C2	
	C3	
	C5	
	C5	
	C6	
	C8	
	C1	9
	C2	
	C3	
	C5	
	C6	
	C8	
	C9	
	C10	
	C11	
Moderate High	C4	1
	C9	5
Extreme Low	C5	1
	C6	
	C11	5
Moderate Low	C1	1
	C10	5

the extreme high proliferation group (Table 2.3). This analysis also revealed a pattern of extreme high proliferation at passage 9 for a majority of the clonal cell lines but not the parental cell lines P0 or P00 (Fig. 2.7 and Table 2.3). Therefore, using the boxplot outlier analysis allowed the identification of subgroups within the Jurkat cell lines at different passages (Fig 2.7).

The outlier analysis shows that a majority (n=9 out of 12) clonal cell lines exhibited extreme high proliferation at passage 9 (Table 2.3), therefore, the distribution of PI for cell lines was graphed per passage (Fig 2.8). This shows that the passage with the highest values are seen at passage 9, relative to passage 1 and 5. Differences in the PI between passages 1, 5 and 9 for all clonal cell lines were found to be of statistical significance using a Friedman Test ($p < 0.1$) (Fig 2.8). This was further analysed using a Wilcoxon Signed-Rank Two-Tailed test as a Post-Hoc test to do a pairwise comparison of two related samples/repeated measurements. Statistical differences between passage 1 and 9 as well as passage 5 and 9 ($p < 0.03$ and $p < 0.003$, respectively) were found (Table 2.4). Overall, this demonstrates that statistical significance between passages were identified. This follows the similar trend seen in earlier results of parental vs clone differences at passage 9 and through variability of PI values amongst specific cell lines to identify differential proliferative groups (Fig 2.7 and Table 2.3).



| Figure 2.8. Boxplot display of proliferative index (PI) data distribution of all clones at passage 1, 5 and 9.

Passages 1, 5 and 9 are shown as blue, red and yellow, respectively. Data points are displayed on the left of the respective boxplot, with a display of the median and mean of PI as indicated by the solid and dashed lines, respectively. Standard deviation of the data per passage is indicated by the peaks of the dashed lines (±). Statistical significance was tested using a Friedman Test, $p < 0.1$. Post-Hoc testing was done using a two-tailed Wilcoxon Signed-Rank test which conferred differences at passages 1 and 9, $p < 0.03$ (*) and passages 5 and 9, $p < 0.003$ (**) (Table 2.4).

| Table 2.4. Post-Hoc test using Wilcoxon-Rank Sum Test results of pairwise-comparisons between passages 1, 5 and 9 (P1, P5, P9, respectively). Significance determined by $p < 0.05$ (*).

p-value	P1	P5
P5	0.91	-
P9	0.03 *	0.003 **

Having characterised the proliferation profiles of parental and clonal cell lines, we next analysed the expression of genes found within the TAL1 CRC with a view to correlating phenotypic profiles with gene expression results.

2.3.5. Optimisation of qPCR Gene Expression assay

SYBR green is a commonly used fluorescent dye for quantifying dsDNA during amplification. However, limitations of non-specific binding must be considered when conducting analysis with this dye (Bustin et al., 2009). This includes primer efficiency analysis of primers used to amplify key cDNA transcripts (Wagner, 2013). This must be done to avoid differences in amplification and misrepresentation of fold change between genes using the $2^{-(ddCT)}$ method (Sreedharan et al., 2018). Therefore, efficiency calculations of individual primer pairs are required prior to gene expression analysis. Primer pairs for *TAL1*, *GATA3*, *RUNX1* and *MYB* were tested for amplification efficiency (Methods: 2.2.9) and all fell between the acceptable efficiency range of 90-110% (Fig 2.9 A-D).

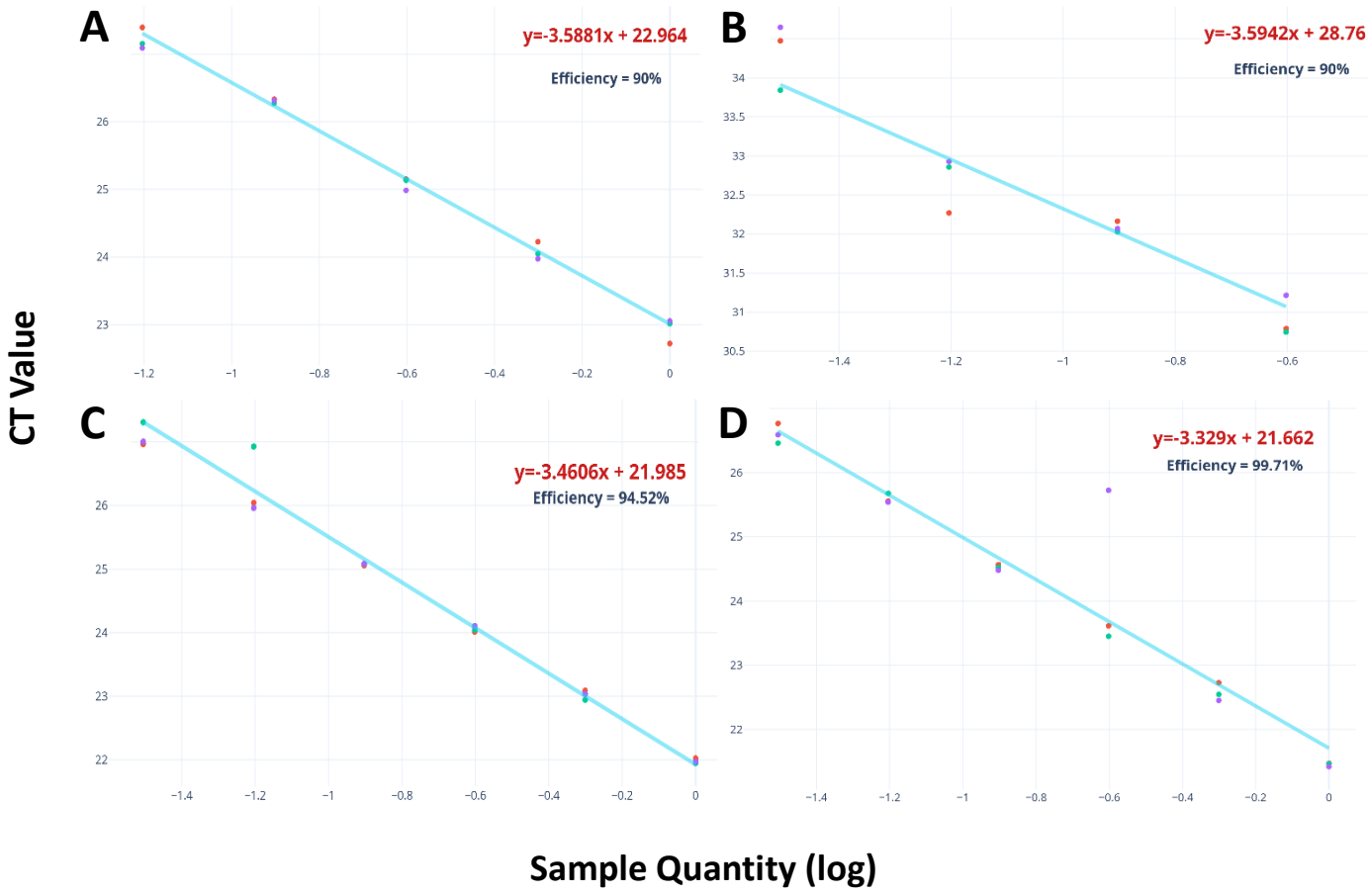
It is also required to assess the 'melting' of template cDNA to assess for specific amplification of one product with the primers of interest (Wagner, 2013). This was done during the primer efficiency qPCR analysis for cycling conditions (Methods: 2.2.10). All primers tested displayed similar melt curves between replicates for each gene tested (Fig 2.10), which displayed a single peak that represented the single melting temperature (T_m) that confirmed a single product being amplified (Bustin and Huggett, 2017).

2.3.6. Clustering of Jurkat clonal cell lines based on gene expression of the TAL1 CRC genes

Data normalised using the $2^{-\Delta\Delta CT}$ were displayed as a Euclidean Distance Hierarchal Clustering Heatmap (Fig. 2.11). Euclidean Distance was used as a measure of correlation-based distances to distinguish patterns of expression amongst all parental and clonal cell lines at all passage points (Glazko and Mushegian, 2010). Glazko and Musehgian (2010) show that Euclidean Distance identifies similarity in gene expression profiles between samples, whereas other correlation-based distance measures over-estimate the divergence of gene expression patterns. This was paired with a hierarchical cluster analysis to further group the patterns of expression amongst all the parental and clonal cell lines (Zhang et al., 2017c). Using this method, three groups can be seen amongst the highlighted black boxes from the clustering analysis (Fig 2.11).

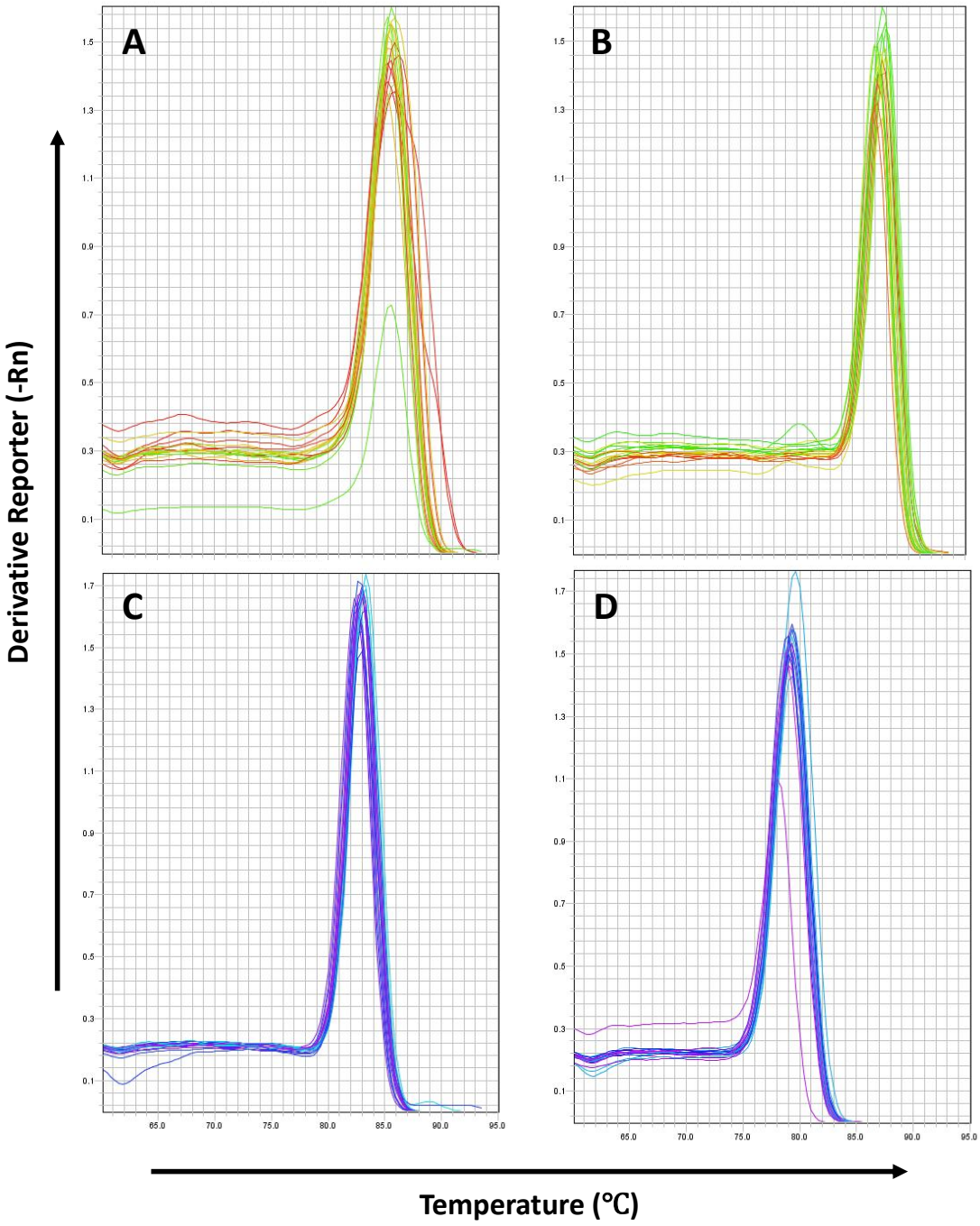
The hierarchical analysis displays three groups that have varying levels of expression of the four T-ALL complex genes, relative to the parental cell line at passage 1 (P0). These groups were classified as *GATA3*$^{+}$, *TAL1*$^{+}$ and *RUNX1/MYB*$^{+}$ based on Euclidean distance that accounts for overall distance amongst cell lines at passage points and gene expression, expressing the relative relationship of these two parameters (Fig 2.11).

The range of standardised expression is between -2 and +2, representing that these differences that are highlighted are across a small range of fold expression (Fig 2.11 – scale bar). For the *GATA3*$^{+}$ group fold change ranged between 0.52 to 2.02 for *GATA3* expression, the *TAL1*$^{+}$ group displayed a range of fold change from 0.39 to 2.02 for *TAL1* expression, and the *RUNX1/MYB*$^{+}$ group displayed a range of fold change from 0.25 and 0.2 to 2.17 and 1.58 for *RUNX1* and *MYB* expression, respectively. Through



| Figure 2.9. (A-D) Primer Efficiencies for TAL1 CRC genes.

Scatter plots of triplicate CT values (y-axis) against log sample quantity (Dilution factor 2) (x-axis) for gene primers, *TAL1*, *GATA3*, *RUNX1* and *MYB*, respectively (10ng of cDNA with a final primer concentration of 0.2µM in reaction). The slope equation of the standard curve (light blue line) is displayed (red) which was used to test the primer efficiency (black) per gene set (Methods 2.2.9).



| Figure 2.10. (A-D) Melt Curve analysis of TAL1 CRC primers.

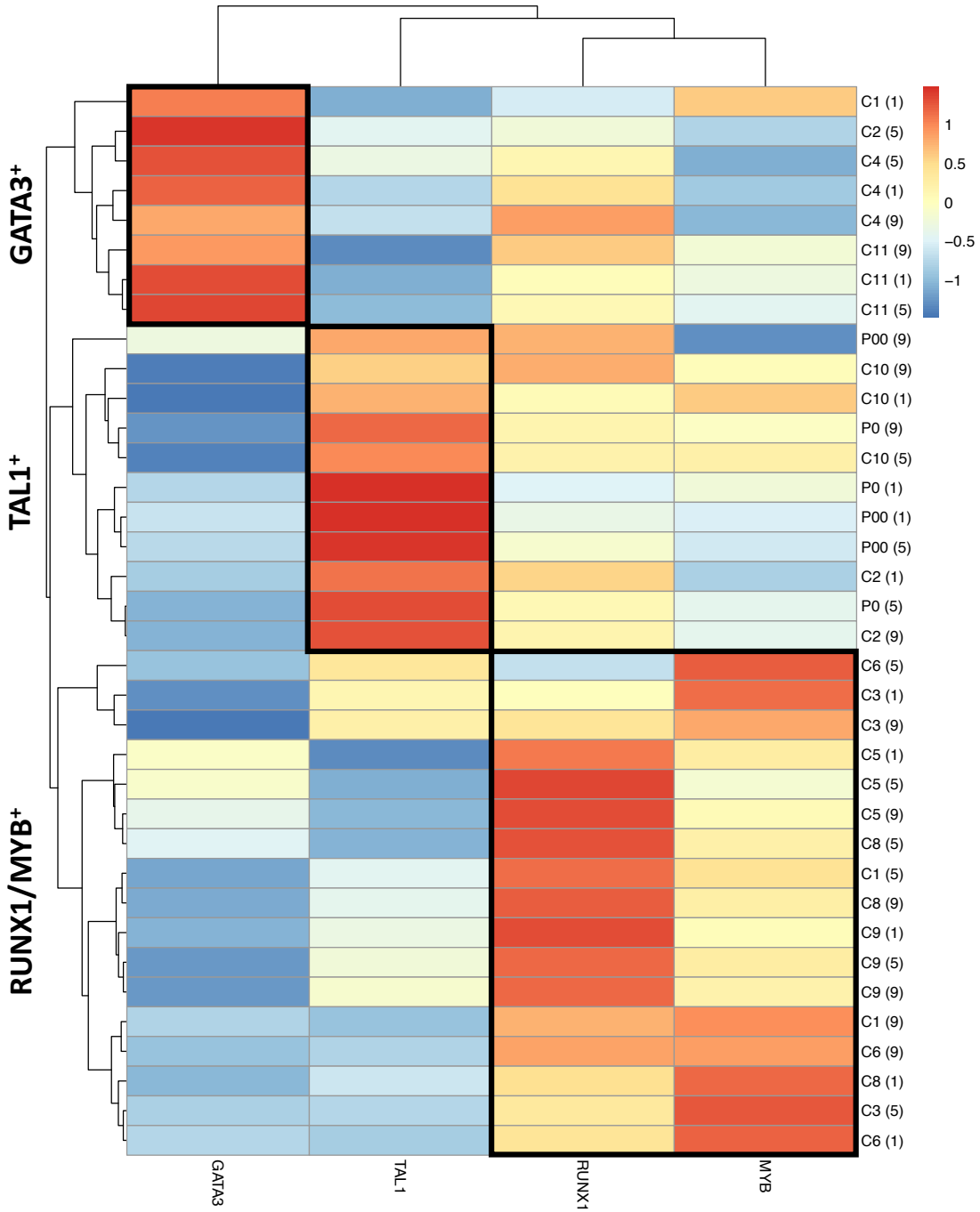
Melt Curve plot of triplicate amplicons generated through primer efficiency optimisation for gene primers (Fig 2.9), *TAL1*, *GATA3*, *RUNX1* and *MYB*, respectively (10ng of cDNA with a final primer concentration of 0.2µM in reaction). Derivative reporter (negative first derivative of normalised fluorescence (Rn) by the reporter) on the x-axis and temperature on the y-axis. A single peak indicates the 'melting' of a single amplicon product after qPCR gene expression analysis.

the use of Euclidean Distance, these fine-tuned patterns of differential expression of relative genes can be seen amongst all Jurkat cell lines and passages. The heatmap further displays an inverse relationship of *TAL1* expression and *GATA3* (Fig 2.11 - indicated by the dark blue regions). However, *RUNX1* and *MYB* expression was moderate (values of 0.5-1) amongst all cell lines that didn't fall within the *RUNX1/MYB*$^{+}$ group regardless of *TAL1* and *GATA3* expression (Fig 2.11).

Based on the Euclidean Distance Hierarchical Clustering heatmap (Fig 2.11), patterns amongst Jurkat cell lines and their expression per passage point revealed sub-groups of Jurkat parental and clonal cell lines (Table 2.5). This displayed a pattern of each Jurkat population falling within the same gene grouping amongst all passages with specifically the parental cell lines grouping within the *TAL1*$^{+}$ group (Table 2.5). However, clones 1 and 2 displayed differential gene expression patterns which fall between different groupings. Specifically, C1 was seen to fall within the *GATA3*$^{+}$ group at passage 1 as opposed to its placement in the *RUNX1/MYB*$^{+}$ group at passage 5 and 9 (Table 2.5). For C2, at passage 1 and 9 it has expression patterns that fall into the TAL1^{+} group as opposed to its placement in the *GATA3*$^{+}$ group at passage 5 (Table 2.5). Individual cell lines at each passage were not tested for differential gene expression as each value represents duplicates and did not provide sufficient data for a robust analysis. Overall, three distinct gene expression groupings were identified and could be used for further analysis into the correlation of these groupings with proliferative ability.

2.3.7. Permutation Analysis for *TAL1* CRC genes

In order to analyse functional properties of differentially expressed genes, a gene set analysis must be done. This is a self-contained method of testing a pre-defined gene



| Figure 2.11. Euclidean Distance Hierarchical clustering of TAL1 CRC gene expression amongst Jurkat cell lines at passages 1, 5 and 9.

Euclidean Distance heatmap of fold expression patterns for parental cell lines and Jurkat cell lines (y-axis) at passages 1, 5 and 9 (brackets) for genes *TAL1*, *GATA3*, *RUNX1* and *MYB* (x-axis) displayed through hierarchical clustering. Fold expression was calculated with the 2^{-ddCT} method relative to the *NFYB* reference gene and the Jurkat parental clonal cell line (P0 at passage 1). Scale of Euclidean Distance is shown from -2 to +2 and hierarchical clustering dendrograms display patterns of Jurkat cell line fold expression of each gene. Black boxes indicate gene groupings of cell lines based on hierarchical clustering.

set for differential expression across two different experimental conditions (such as parental cell line vs clonal cell line) (Chang and Tian, 2016).

A q-value output is utilised as it measures level of significance in regards to a more stringent false discovery rate method rather than the p value false positive rate (Table 2.6) (Storey and Tibshirani, 2003). The results of the permutation analysis display no statistical differences between gene sets when comparing Jurkat parental (P0) at each passage (Control) and the remaining cell lines at all passages and amongst each passage (Table 2.6). It is only at passage 5 (*) that a statistical difference between the *TAL1* gene set can be seen to be significantly up-regulated in the control (P0 and passage 5) ($q < 0.1$ – threshold determined by GSALightning developers) (Table 2.6). This suggests that within passage 5, *TAL1* gene expression is significantly higher within the parental cell line, P0, in comparison to the remaining clonal cell lines.

2.3.8. Differences between Jurkat clonal cell line gene expression between passages

Based on the permutation analysis, further investigation into the differences within gene sets between passages was instigated. This was done using a Friedman Test as an assessment of multiple sets non-parametric data representing different analysis at different passage points (Eisinga et al., 2017). This concluded that all genes tested amongst the cell lines (n=12) were differentially expressed between passages ($p < 0.05$ for *TAL1* and *MYB*, $p < 0.01$ for *GATA3* and *RUNX1*) (Appendix – Supp. Table 7.8). As a result of this, a Wilcoxon Signed-Rank Test was done for all comparisons as a Post-Hoc test (P1 vs 5, P1 vs 9 and P5 vs 9) (Appendix – Supp. Table 7.9). The Wilcoxon test concluded that significant differences within *TAL1* and *GATA3* gene expression was seen for passages 1 and 5 and 9 and 5 ($p < 0.05$). However, for expression of *RUNX1*, passages 1 and 5 and 1 and 9 were statistically different ($p < 0.05$) (Appendix – Supp. Table 7.9). For the expression of *MYB*, passages 1 and 5 were the only comparisons that were of significance ($p < 0.05$) (Appendix – Supp. Table 7.9). Therefore, for genes

| Table 2.5. List of Jurkat cell lines and their passage number for the gene expression groups identified through Euclidean Distance and Hierarchical Clustering of fold expression patterns.

Group	Cell Line	Passage
GATA3^{+}	C1	1
	C2	5
	C4	1
		5
		9
	C11	1
		5
		9
TAL1^{+}	P0	1
		5
		9
	P00	1
		5
		9
	C2	1
		9
	C10	1
		5
		9
RUNX1/MYB^{+}	C1	5
		9
	C3	1
		5
		9
	C5	1
		5
		9
	C6	1
		5
		9
	C8	1
		5
		9
	C9	1
		5
		9

| Table 2.6. Permutation q-values amongst gene sets at passage 1, 5, 9 and all passages combined determined by GSALightning R package.

q-value	Passage 1		Passage 5		Passage 9		All Passages	
Gene	Control	Experiment	Control	Experiment	Control	Experiment	Control	Experiment
GATA3	0.16	0.85	0.43	0.99	0.56	1	0.50	0.88
MYB	0.16	0.92	0.28	0.99	0.35	1	0.50	0.88
RUNX1	0.16	0.83	0.27	0.99	0.37	1	0.59	0.88
TAL1	0.16	1	0.03*	0.99	0.42	1	0.17	0.96

[1] Analysis based on the upregulation of the control (P0) or experiment (all clones). Significance at $p < 0.1$ (*).

TAL1 and *GATA3*, there is a progression of change between passages 1 and 5 and 5 and 9, however the difference is not large enough between passage 1 and 9 to be detected. Whereas *RUNX1* showed changes of expression from passage 5 and onwards. For the expression of *MYB*, changes are seen between passages 1 and 5, however are reverted back to similar distributions of expression after passage 5 according to the Wilcoxon-Rank Sum test (Appendix – Supp. Table 7.9). This shows that differential expression of each of these genes is observed between passages.

2.3.9. Statistical testing of Jurkat gene expression groups

Gene expression groups identified through the Euclidean Distance Hierarchical Clustering were statistically tested for differences between expression of each gene. This was conducted using a Kruskal-Wallis test as a non-parametric assessment of multiple samples (Nahm, 2016) (Table 2.7). The initial Kruskal-Wallis test will identify differences between at least two samples and a further Post-Hoc test must be done for pairwise comparisons (Kim, 2014). This found statistical differences between groups with *TAL1* and *GATA3* expression ($p < 0.03$ and 0.02, respectively), however not found for *RUNX1* and *MYB* (Table 2.7). A Post-Hoc test was conducted to further adjust p-values for multiple comparisons. This was done with the Dunn test and further

adjusted using the Holm FWER and the Benjamini-Hochberg FDR test. Both adjustments concluded that no statistical differences between the expression of *TAL1* and *GATA3* amongst the expression groups were found (Appendix – Supp. Table 7.10-7.13).

Due to no statistical differences being found for gene expression groups *TAL1*$^{+}$, *GATA3*$^{+}$ and *RUNX1/MYB*$^{+}$, a correlation test was conducted. A Spearman's Rho assessment was done for the genes expressed amongst each expression group. The Spearman's Rho test accounts for non-parametric data and develops a correlation coefficient based on the calculated ranks from the data tested (Akoglu, 2018). It was determined that all groups and all samples tested had gene expression correlation amongst all genes (Appendix – Supp. Table 7.14 - 7.17).

| Table 2.7. Kruskal-Wallis test statistics of differences between gene expression groups (TAL1^{+}, GATA3^{+} and RUNX1/MYB^{+}) for expression of the genes *TAL1*, *GATA3*, *RUNX1* and *MYB*.

Gene Expression	p-value < 0.05	Significance
TAL1	0.03	Yes
GATA3	0.02	Yes
RUNX1	0.94	No
MYB	0.61	No

2.3.10. Gene Expression and its relationship with the Jurkat cell line proliferation phenotypes

An understanding of the phenotypic differences identified within Jurkat cell lines and the gene expression profiles of these cell lines is required to understand heterogeneity. Jurkat cell lines were characterised by their proliferative ability as a measure of a proliferation index (PI) – an average number of cell divisions of cells 'responding' to the re-introduction of serum after starvation. These cell lines were categorised into groups determined by their proliferative ability and a boxplot analysis of outliers. The

groups identified were the moderate and extreme high and low proliferative groups, where a majority of cell lines fell within the extreme high category (n = 20 out of 27).

The cell lines within these groupings were then correlated to the expression profiles of the tested genes. This was conducted with the Spearman's Rho Test for non-parametric data. It was found that there was no correlation of the relative fold gene expression of the T-ALL genes to the proliferative abilities of the cells within the extreme high proliferation group ($p > 0.05$) (Appendix – Supp. Table 7.18). Analysis could not be done with the moderate high, moderate low and extreme low categories as their sample sizes ranged from n=4-6. Therefore, no relationship between proliferation and proliferative abilities of the moderate high, extreme high, moderate low and extreme low proliferative groups were found. A correlation test between the gene expression groups (*TAL1*$^+$, *GATA3*$^+$ and *RUNX1/MYB*$^+$) and proliferation was also conducted, which yielded no correlations (Appendix – Supp. Table 7.19). Overall, demonstrating that TAL1 CRC genes tested had no correlation to the proliferation of Jurkat cell lines.

2.4. Conclusion

In conclusion, the aim of this analysis was set out to identify differences between Jurkat clonal populations and establish groupings of these cell lines by their proliferative ability using a cell proliferation CFSE assay and were correlated to gene expression profiles by qPCR with relative expression to the parental Jurkat population (P0) at passage 1. It was found that proliferation (as measured by PI) between parental populations (P0 and P00) and the derived clonal cell lines was statistically different at passage 9. The qPCR gene expression analysis identified key gene identity groups (*TAL1*$^+$, *GATA3*$^+$ and *RUNX1/MYB*$^+$) that suggest differential gene expression patterns between cell lines. However, despite these patterns being revealed through Euclidean Distance Hierarchical Clustering, no statistical differences between these groups

could be distinguished. Finally, correlations between proliferation index values acquired from phenotype characterising of Jurkat cell lines and gene expression profiles for these cell lines was conducted. This concluded no correlations between gene expression groups and proliferation groups amongst Jurkat cell lines. Therefore, these patterns suggest a deeper-rooted mechanism, or a larger scope of genes that may be involved or reflect Jurkat proliferative abilities.

Chapter 3 - Bioinformatic analysis of the *TAL1* locus

3.1. Introduction

Bioinformatics is an interdisciplinary field at the intersection of molecular biology, genetics, computer science and statistics (Can, 2014). It investigates complex biological problems using large data sets by using computational algorithms and models (Can, 2014). The potential power of genomics in a clinical health setting is exemplified by the Broad Cancer Genome atlas, as well as the UK Biobank (Sudlow et al., 2015; Tomczak et al., 2015), with initiatives such as these providing a detailed molecular picture of our genomes with a view to translating genomics into better healthcare. Through advancements in next-generation sequencing technology and robust data sets created by the ENCODE project, bioinformatics pipelines are becoming increasingly easy to use and understand (Coccaro et al., 2019). Not only does bioinformatics analysis allow for robust high-throughput processing of such complex data but, there is a community of practice that requires authors to make their data public through submission to open and accessible databases, such as the NCBI GEO datasets (Clough and Barrett, 2016). This allows for analysis of the data by scientists who work within the field but may not have generated the data themselves. The involvement of a larger community in data analysis has benefits in terms of quality control of the data as well as novel analyses that arise from the combination of multiple data sets.

Next-generation DNA sequencing of DNA recovered from chromatin immunoprecipitation experiments (ChIP-Seq) maps the genome-wide locations of transcription factor (TF) binding and the sites of histone modifications (Jiang and Mortazavi, 2018). The ENCODE project has provided a valuable archive of data that

maps epigenetic markers throughout the genome in multiple cell types (Consortium, 2004; Davis et al., 2018; ENCODE Project Consortium, 2011; Zacher et al., 2017). These epigenetic markers identify regulatory DNA elements, such as enhancers, that modulate gene expression (Blinka et al., 2017). To map putative enhancers requires information on specific chromatin markers such as histone subunit 3 lysine 27 acetylation (H3K27ac) and DNAseI Hypersensitivity (DHS) (Calo and Wysocka, 2013). Promoters can be differentiated from enhancers through the additional enrichment of histone subunit 3 lysine 4 trimethylation (H3K4me3) in combination with H3K27ac and DHS (Calo and Wysocka, 2013).

Within Jurkat T-ALL cells, Mansour et al. (2014) have discovered a key functional super-enhancer region that is enriched in read depth and breadth for the active enhancer marker, H3K27ac. It was revealed that this region contains a heterozygous 12-bp insertion that aligns with TAL1, GATA3, RUNX1 and MYB ChIP-seq enrichment (Mansour et al., 2014). This mono-allelic overexpression of TAL1 in Jurkat T-ALL cells is also further correlated with DNA hypomethylation at the *TAL1* TSS, allowing for expression of the gene, however DNA methylation across the locus at intergenic and intragenic regions that map to regulatory elements in the context of clonal heterogeneity is not known (Haider et al., 2018). Although the 12-bp insertion mutation has created a super-enhancer, we suggest that other intra- or intergenic enhancers may exist across the *TAL1* locus.

The aims of this chapter are to use bioinformatic analysis of large genomic datasets to identify intra- and intergenic enhancers across the *TAL1* locus and to then determine the DNA methylation patterns in the parental and clonal Jurkat T-cell lines (Chapter 2) at possible intra- and intra-genic enhancers in Chapter 4.

3.2. Methods

3.2.1. Source and type of ChIP-seq Data

ChIP-seq BAM files (aligned reads) and BroadPeak/NarrowPeak files (Peak Data) were downloaded for markers of DHS, H3K27ac and H3K4me3 (BroadPeak files were prioritised over NarrowPeak files because histone modifications are better analysed as broad domains of enrichment (Starmer and Magnuson, 2016)) from the ENCODE Project Consortium, which provides high-quality data available in file formats that include processed and aligned raw sequencing data (https://www.encodeproject.org/). ChIP-seq data for Jurkat and T-cell primary cell lines for H3K27ac and H3K4me3 were downloaded from the European Nucleotide Archive from The European Bioinformatics Institute (https://www.ebi.ac.uk/ena) for FASTQ files. ChIP-Seq data included: Jurkat cell line, a B-cell line (CD20RO01794 and CD20RO01778), monocyte primary cells (MonoCD14$^+$), normal CD4$^+$ T-cells (primary T-Helper (Th) 1, 2 and 0) and other cell lines that aren't categorised in these groups such as A549 (small lung carcinoma), HMEC (endothelial mammary) and NHEK (Keratinocytes). Origins and accession numbers for all files downloaded are in Appendix – Supp. Table 7.20 – 7.23.

3.2.2. Quality Control Processing of ChIP-Seq Files

All BAM and FASTQ files were run through the 'FastQC' program (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). The FastQC program is one of the most commonly used quality control programs for FASTQ, SAM and BAM files that provides metrics on the quality of the sequencing run per sample (Park et al., 2017; Trivedi et al., 2014). One of the metrics that can be derived from the FastQC program is a Phred Score. Phred scores determine the quality of base-calling during sequencing and thus has been used as a measure for the files processed for this project.

Phred scoring of data is essential for ensuring quality for bioinformatics analysis, as processing of low-quality files will result in inaccurate outputs for analysis (Zhang et al., 2017b). It is recommended that files with a median Phred score below 25 or 20 are flagged with a warning or a failure for quality, respectively (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). This measure was used to assess files tested and generated metrics from FastQC of the overall quality and for specific categories that included: sequencing adaptor content and the need to trim sequencing files, GC content, and sequence length distribution (example of quality score across all bases in Appendix - Supp Fig 7.2).

In cases where sequencing files displayed a Phred score below 25 (but not below 20), trimming was conducted to remove low quality reads and remaining adaptors. The package 'FastP' was utitlised for its ability to provide pre-processed clean data for ChIP-seq analysis (Chen et al., 2018) (command line in Appendix – 7.2.3.a). The FastP program also provides metrics on quality of the files before and after trimming (Supp. Table 7.24). After trimming, files were run through 'FastQC' again to reassess quality, if they still fell below the threshold of a Phred score of 25, they were eliminated from further analysis (Supp. Table 7.24).

3.2.3. FASTQ sequencing file genome alignment, conversion and peak calling

After testing the quality of FASTQ files downloaded from the ENA database, FASTQ files were aligned to a reference genome (GRCh37/hg19) using the Python package, 'Bowtie2' (Bailey et al., 2013; Zhang et al., 2016b). The resulting SAM files were converted using the default parameters for single end read files and further converted to a BAM file using the 'Samtools' package (Li et al., 2009; Steinhauser et al., 2016). Examples of command lines for this processing pipeline can be found in Appendix (7.2.3.b and c).

Quality checks of SAM and BAM files can be done using the FastQC program to ensure that processing milestones within the pipeline are conducted properly. These were done for the remaining files left for the analysis. When these SAM/BAM files are generated through the processing pipeline, they are unsorted and require an additional step to have them sorted by coordinate (Li et al., 2009). The 'Samtools' package also conducts sorting of converted BAM files (command line in Appendix – 7.2.3.c).

Once all BAM files from ENA were processed, the remaining files needed for differential binding analysis required calling of ChIP-seq read peaks. Using the online database 'Galaxy' (https://usegalaxy.org/), all BAM files were uploaded, including control files from designated experiments, and were called for peaks using the 'MACS2 peakcall' option. The following default parameters for MACS2 were used: Building of a shifting model, lower mdfold bound: 5, upper mfold bound: 50, band width: 300, peak detection based on: q-value, FDR threshold: 0.05. The output provided a quality assessment through a peak model plot of the aligned datasets and cross-correlation plot as well as provide BroadPeak files and a BED file of the peaks called.

Peak model plots and the cross-correlation plots provided by MACS2, display a visual representation of the quality of the ChIP-Seq data. Within the cross-correlation plot, a single peak should appear displaying a larger fragment-length peak compared with a read-length peak (see: https://hbctraining.github.io/Intro-to-ChIPseq/lessons/06_combine_chipQC_and_metrics.html) (Appendix - Supp. Fig 7.3A). Through the output of a peak model plot by MACS2, further verification of the quality of the ChIP-seq files could be seen through the combination of positional information and reads from both strands to more accurately map the location of a

peak. (see: https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html) (Appendix - Supp. Fig 7.3B).

3.2.4. ChIP-Seq Differential Binding Analysis (DBA)

A common issue found with the analysis of ChIP-seq data is the comparison and normalisation of the total sequencing library and peak analysis between different experiments (Wu et al., 2015). Typically, input controls are used to account for background noise, however, regions with extreme copy number variation could still impact the final result (Wu et al., 2015).

The R programming package 'DiffBind' was utilised for its full library normalisation and its use of the known ChIP-seq differential binding programs 'DESeq2' and 'edgeR' (Brown, 2011; Steinhauser et al., 2016). As a first step in identifying differentially bound peaks between different samples, DiffBind was used to assess the quality of replicate samples by assessing the degree of consensus of identified peaks, as FastQC does not assess the quality of the experiment prior to sequencing (biological replicate consensus). We set a minimum threshold of 70% peak consensus between replicate samples. However, an exception was made for data from the Jurkat cell line (Table 3.1).

Next, the correlation between samples were analysed through counting the reads within the data which was normalised using the TMM method (Trimmed mean of M). This normalises two samples by detecting overall enrichment between loci without the assumption of differentially expressed reads (Robinson and Oshlack, 2010). Parameters were set to have a minimum overlap of 2 samples with consensus peaks (see: https://support.bioconductor.org/p/57809/). After normalisation, read count heatmaps that display relationships between samples were produced with the inclusion of a z-score plot to show the standard deviation distribution of data

according to the population mean (e.g. Fig 3.2) (Curtis et al., 2016). Then, a differential binding analysis (DBA) was conducted to identify contrasts between various groups established, such as Lymphocyte vs Other Non-Immune cell lines (H3K27ac and H3K4me3) OR Immune vs Non-Immune cell lines (DHS), T-cell vs Jurkat and Jurkat vs DND41 (command line in Appendix – 7.2.3.d). This outputs a DBA heatmap that displays the relationship between samples that were contrasted and a DBA report of the Genomic Ranges (GRanges Object). The GRanges report displays the regions of differential binding, the fold change between contrasts, raw p-values and corrected FDR values found between contrasted regions (example of Top 50 results for DHS DBA – Supp. Table 7.25).

Using 'DiffBind' the following assessments of differential binding were conducted: immune cells and cell lines (Jurkat, primary cell lines, Th1, Th2, Th0 and MonoCD14) in comparison to other non-immune cell lines (SAEC, NHEK, A549, HMEC and 8988t) to establish a differential binding signature of immune cells overall for DHS DBA, or non-immune cell lines in comparison to lymphocyte cell lines (B-cells CD20RO01778 and CD20RO01794 and Jurkat) and primary cells (primary Th1, Th2 and Th0) for H3K27ac and H3K4me3 DBA. Lastly, a comparison between other primary CD4+ T-cell lymphocytes (primary Th1 and Th2) and the Jurkat T-ALL cell line was conducted to establish differential binding specific to TAL1+ T-ALL. To further this analysis, ENCODE H3K4me3 and DHS data from the T-ALL cell line DND41 was used to identify differential binding against the Jurkat cell line (T-ALL vs TAL1-T-ALL) (Accession Numbers in Appendix – Supp Tables 7.20 – 7.23).

3.2.5. Processed data visualisation in the UCSC genome browser

Pipeline processed ChIP-seq files from ENA and the pre-processed files from ENCODE were then visualised within the UCSC genome browser (https://genome.ucsc.edu/). Non-ENCODE BAM files were converted to BIGWIG

format through the database 'Galaxy' (https://usegalaxy.org/), using the 'bamCompare' option from 'deepTools'. This allowed for the normalisation of BAM files to input control BAM files with the output of a BIGWIG file for visualisation in the UCSC genome browser (Ramírez et al., 2014). These files were then uploaded into the 'CyVerse' server to provide a URL of the BIGWIG files to upload into the UCSC genome browser.

All files for regulatory DNA element markers (H3K27ac, DHS and H3K4me3) for different cell lines were uploaded into the UCSC genome browser, as well as a file of regions identified through DiffBind that included statistically significant differentially bound peaks between cell lines tested. Looping interactions between regulatory elements across the *TAL1* locus was also visualised through the GeneHancer database. GeneHancer is a database that provides data related to enhancers and frames regions of enhancer interactions throughout the genome. It specifically integrates data from a total of 434, 000 reported enhancers from the ENCODE project, the Ensembl regulatory build, VISTA Enhancer browser and the Functional Annotation of the Mammalian genome (FANTOM) project (Fishilevich et al., 2017). This database links enhancers to genes by using information such as tissue expression correlation between genes and eRNAs, transcription factor binding sites, Hi-C data, and expression quantitative trait loci variants (eQTLS) and was therefore used for this analysis. (Fishilevich et al., 2017). With an additional track from the GTEx database of RNA transcripts for thyroid and whole blood cell types.

| Table 3.1. List of cell lines tested for consensus peaks in replicate samples using 'DiffBind' consensus peak overlap.

Primary cells/Cell Line	Marker	Tissue	Rep 1	Consensus	Rep 2	Total Peaks	Percentage of replicate peak overlap[1]
CD20RO01778	H3K4me3	B-Cell	21177	35410	20778	77365	45.7
DND41	H3K4me3	T-ALL		25790		25790	100
Jurkat	H3K4me3	TAL1+ T-ALL	15573	31423	27187	74183	42.4
MCF7	H3K4me3	Adenocarcinoma	61396	47209	43349	151954	31.1
SAEC	H3K4me3	Lung Epithelial	12026	24776	15439	52241	47.4
NHEK	H3K4me3	Keratinocytes	13413	23971	16045	53429	44.9
A549	H3K4me3	Lung Cancer		74621		74621	100
A549v2	H3K4me3	Lung Cancer	23520	39428	32169	95117	41.5
HMEC	H3K4me3	Endothelial Mammary Gland	29578	33548	29484	92610	36.2
AG04449	H3K4me3	Skin Fibroblast	32524	40704	26667	99895	40.7
AG04450	H3K4me3	Lung Fibroblast	40488	30358	26667	97513	31.1
DND41	H3K27ac	T-ALL		39585		39585	100
HMEC	H3K27ac	Endothelial Mammary Gland		57262		57262	100
Jurkat	H3K27ac	TAL1+ T-ALL	1504	9195	4485	15184	60.6
NHEK	H3K27ac	Keratinocytes		57797		57797	100
Th2	H3K27ac	T-Cell	791	32772	744	34307	95.5
A549	DGF	Lung Cancer	23273	95564	39452	158289	60.4
CD20RO01778	DGF	B-Cell	57357	89317	104368	251042	35.6
CD20RO01794	DGF	B-Cell		116032		116032	100
SAEC	DGF	Normal Lung	42682	155695	17874	216251	72.0
8988t	DGF	Pancreatic Cancer		144165		144165	100
Th1v2	DGF	Th1		351905		351905	100
MonoCD14	DGF	Monocytes		159457		159457	100
Jurkat	DGF	TAL1+ T-ALL	86560	136177	142477	365214	37.3

[1]Cell lines with a percentage overlap of > 70% (green) were used in the following differential binding analysis whereas replicate consensus < 70% were excluded (red). Jurkat cell lines were used regardless of percentage overlap (yellow).

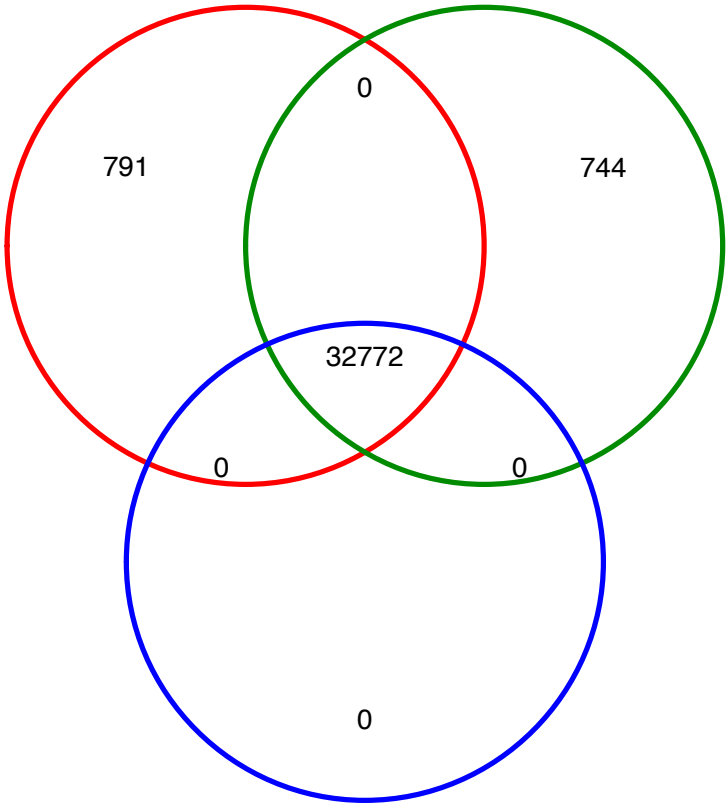
3.3. Results

3.3.1. Differential Binding Analysis of ChIP-seq data between multiple cell lines

The isolation and characterisation of Jurkat T-ALL clonal cell lines relative to a parental cell line showed increased proliferation at high passage numbers, as well as a decrease in the expression of *TAL1* (Chapter 2 - Fig 2.6 and 2.11, respectively). It is possible such differences might reflect genetic and/or epigenetic differences between these cell lines at the *TAL1* locus. In preparation for an analysis of DNA methylation and genetic variation across the *TAL1* locus, we used a bioinformatic analysis to identify putative regulatory elements across the *TAL1* locus in Jurkat T-ALL cells.

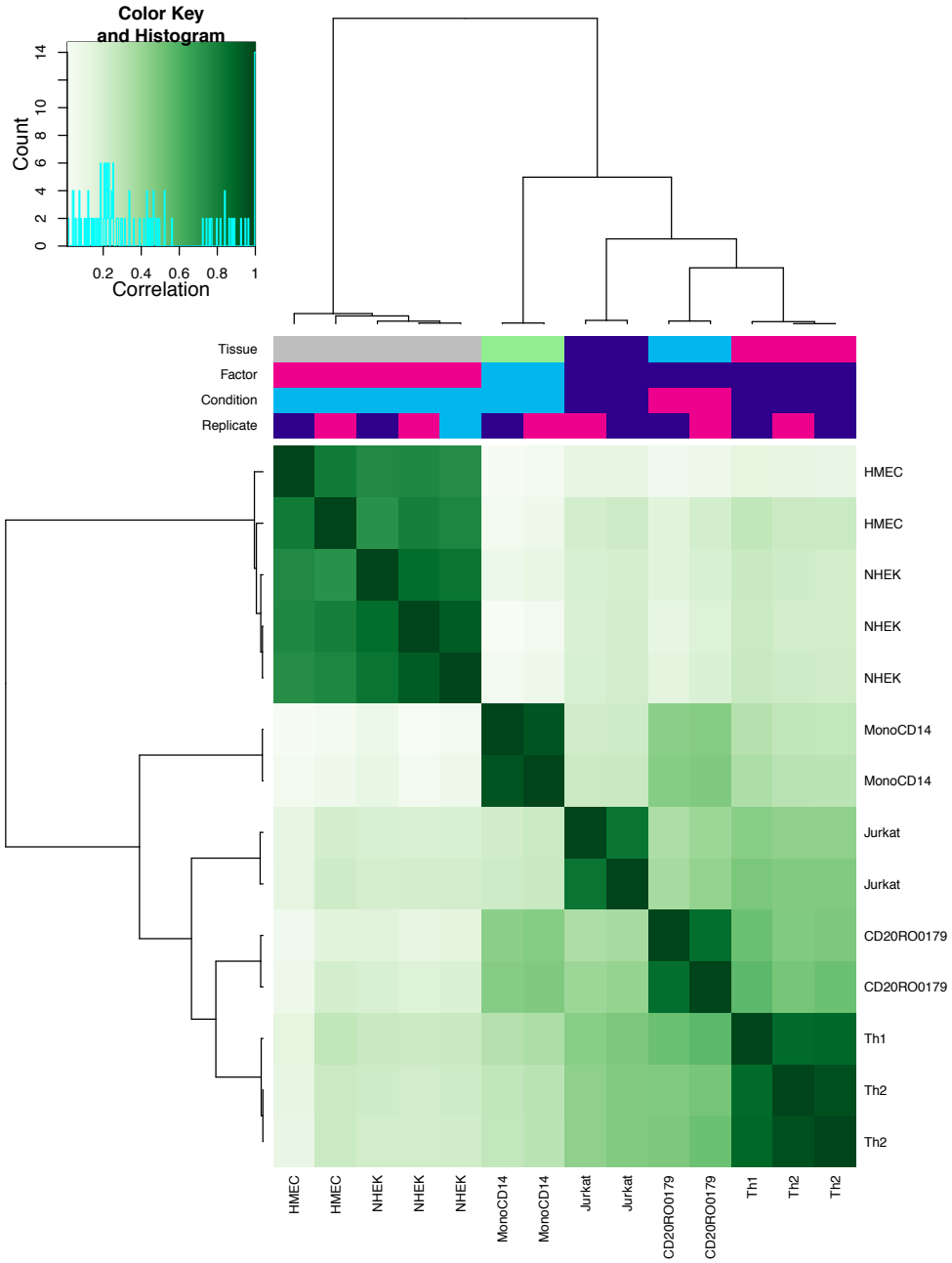
The combination of DHS, H3K27ac and H3K4me3 provides a map of open chromatin and active enhancers and promoters. Using these markers, regulatory elements were mapped across the *TAL1* locus and were tested for differential binding between cell groups stated (Methods: 3.2.4). Firstly, to ensure replicates utilised for the analysis were comparable for analysis, they were tested for overlap of consensus peaks with a threshold of 70%. This is to ensure that differences between replicates would not influence the DBA between cell groups tested. For example, consensus of Th2 cell replicates displayed that replicate 1 had 791 different peaks (red circle), replicate 2 had 744 different peaks (green circle), with 32772 overlapping peaks (blue circle), corresponding to a consensus peak overlap of 95% (Fig 3.1). This was done for all cell lines tested for each of the regulatory element markers tested. Cell lines with less than 70% overlap between replicates were excluded from the analysis, except for the Jurkat cell line (Table 3.1).

After confirmation of consensus peaks between replicates, a normalised read count of peaks and a measure of binding affinity (a measure of read depth at peaks) for each



| Figure 3.1. Venn diagram of H3K27ac enriched peaks within Th2 cell line replicates (red and green circles).

'DiffBind' generates a third data set of the overlapping consensus peaks found between the two replicates (blue circle). The percentage of peaks overlapped amongst replicates as a quality control for the differential binding analysis was taken from this consensus peak analysis. This example displays a consensus overlap of 95%, therefore is an example of replicate data used for the analysis.

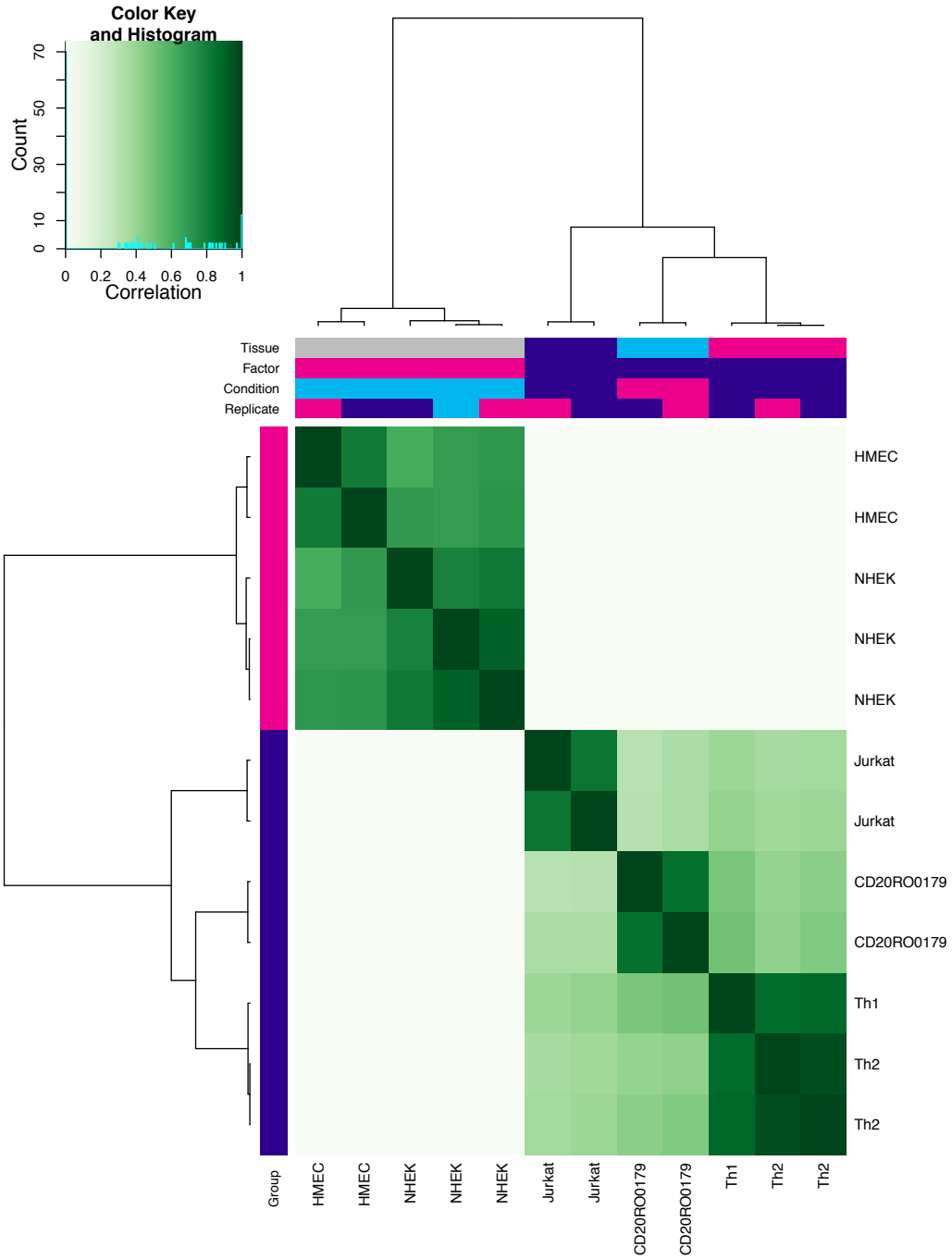


| Figure 3.2. Correlation heatmap based on read counts performed by the R package, DiffBind for H3K27ac ChIP-seq data.

Cell line replicates are clustered based on read counts of H3K27ac and relationships are displayed with dendrograms, along with the tissue type, factor type (other, immune or lymphocyte), condition (other, T-cell or B-cell) and replicate number. The distribution of read counts are displayed as a z-score correlation histogram (top left). The clustering of non-immune cell lines (HMEC and NHEK) can be seen relative to immune cell types (MonoCD14, Jurkat, CD20RO01794, Th1 and Th2), with further clustering of primary T-helper cells relative to monocytes and the Jurkat cell line.

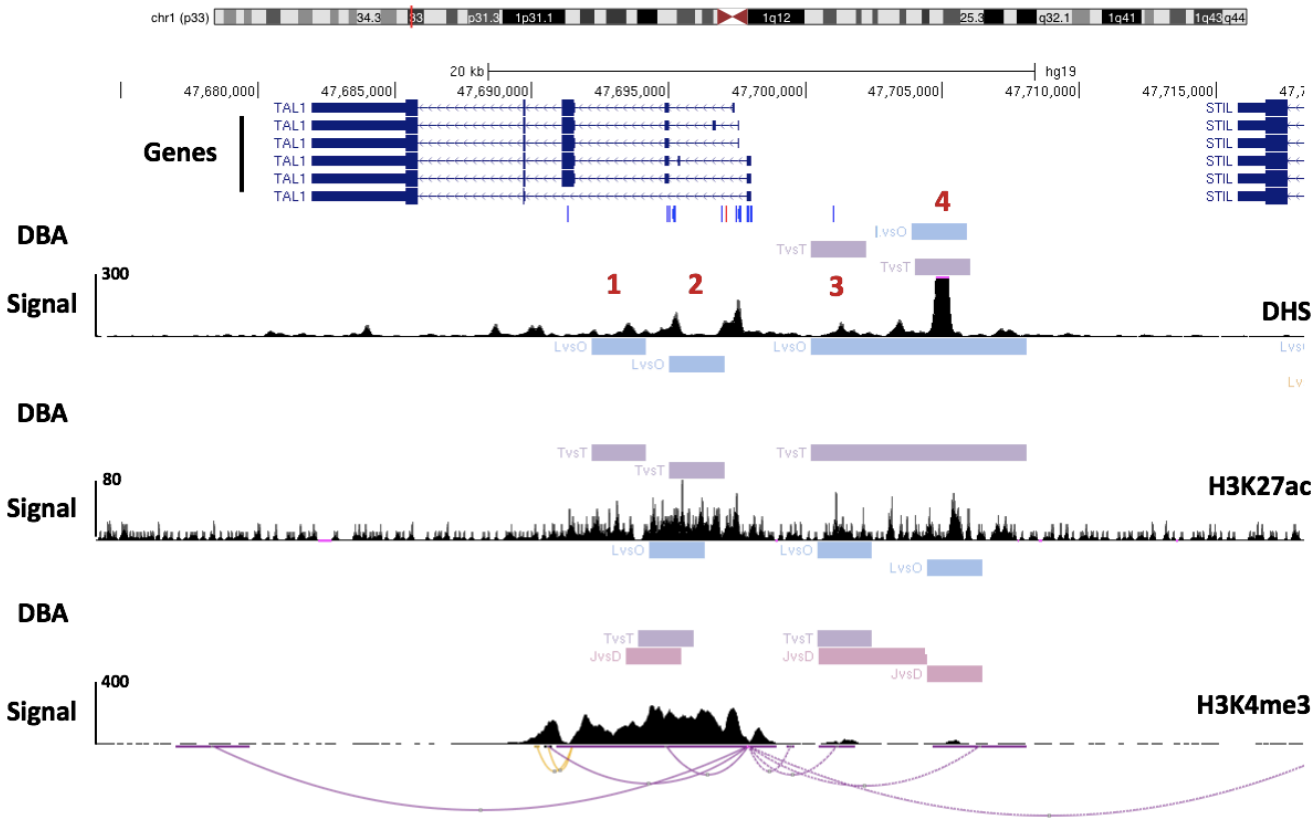
marker tested was conducted. This predicts that the clustering of cell lines will reflect their origin and is displayed in a correlation heatmap (Fig 3.2). As an example, for H3K27ac, a correlation heatmap was produced to display the relationship between samples within the groups designated above as well as a histogram plot of the z-score distribution of the data analysed (Fig 3.2 – top left). It is seen that cell types are grouped within the correlation heatmap: non-immune cell lines, HMEC and NHEK in the top left (Fig 3.2) and that CD4+ T-cells cluster at the right bottom corner (Fig. 3.2). All immune cells (MonoCD14, Jurkat and CDRO01794) are identified as unique but fall into a larger "immune cell" clade when compared to the non-immune cell lines (Fig 3.2 – bottom right). This showed that immune cells have a greater identity in terms of shared peaks when compared to non-immune cell lines (Fig 3.2 – top left). Thus, this analysis showed that cells were clustered in a way that reflected epithelial/non-immune and immune cell samples used in this study. This analysis allowed us to investigate differential binding affinity (DBA) between and within cell types.

A DBA was conducted for all cell groups to be tested, DiffBind generates a Granges Object of differential binding sites between cell groups throughout the entire genome. Conducting DBA resulted in over 20,000 differentially bound sites for DHS, H3K27ac and H3K4me3, except for the analysis between immune cells and cell lines in comparison to other non-immune cell lines for DHS (38 results) (data not shown). Firstly, a DBA contrast (threshold FDR 0.05) between lymphocyte cell types (primary CD4+ T-cells, B-cell line and Jurkat) and non-immune cell types (as stated in Fig 3.2) was conducted (Fig 3.3). This showed the contrast between these cell types and are grouped accordingly in the heatmap (Fig 3.3 – top left and bottom right). The absence of data within the heatmap (top right and bottom left) indicates an absence of shared differential binding sites between the two groups tested (Fig 3.3). This was also done for the Jurkat cell line (T-ALL) and primary CD4+ T-cells and the Jurkat cell line and



| Figure 3.3. Differential binding analysis (DBA) heatmap based on contrasts of H3K27ac enrichment between lymphocyte cell lines (Jurkat, CD20RO01794, Th1 and Th2) and non-immune cell lines (HMEC and NHEK).

Cell line replicates are clustered based on differential binding of H3K27ac and displayed as a dendrogram (y-axis). A dendrogram also displays relationships of tissue type, factor, condition and replicate numbers amongst samples tested. The distribution of the data is displayed as a z-score correlation histogram (top left). The contrast analysis has further highlighted trends seen in Fig 3.2, but also displays no shared differential binding between immune cell types and other non-immune cell types.



| Figure 3.4. UCSC Genome browser display of the regulatory element markers, DHS, H3K27ac and H3K4me3 across the *TAL1* locus and with localising regions of differential binding.

Differential binding analysis was conducted between immune primary cells and cell lines (primary MonoCD14, primary Th1 and Th0, Jurkat and B-cell cell line, CD20RO01794) and non-immune cell lines (8988T and SAEC) (I v O - blue) for DHS **or** lymphocyte cells and cell lines (primary Th1 and 2 cells, Jurkat and B-cell cell line, CD20RO01794 (H3K27ac) or CD20RO01778 (H3K4me3) and non-immune cell lines (NHEK and HMEC for H3K27ac and MCF7 and SAEC for H3K4me3) (L v O - blue). As well as comparisons of primary CD4+ T-cells (Th1 and Th0 for DHS, Th1 and Th2 for H3K27ac and H3K4me3) and the Jurkat cell line (T vs T - purple) and the Jurkat cell line and the DND41 (T-ALL) cell line (J vs D – pink) (Table 3.2). Regions identified from the DBA within the *TAL1* locus were plotted as boxes above each designated regulatory element marker to show the site and breadth of differential binding between groups tested. The ideogram of chromosome 1 displays the location of the *TAL1* locus along the chromosome (red line) and the isoforms of Ref-Seq genes present within the locus (*TAL1* and *STIL*) are displayed. Regulatory element markers (DHS, H3K27ac and H3K4me3) display read-depth on the y-axis. GeneHancer interactions are displayed as loops (bottom panel) for regulatory element interactions. FANTOM TSS Peaks displayed show sites of transcriptional activity for forward and reverse strands (red: forward, blue: reverse). Regions are indicated by red numbers for each regulatory element marker. Red numbers indicate common regions of differential binding between regulatory element markers that map to predicted downstream TSS, H3K27ac and H3K4me3 enriched region (1), TSS, H3K27ac and H3K4me3 peak (2), predicted downstream intergenic enhancer from *MuTE* insertion (3) and the *MuTE* insertion enhancer peak (4).

the T-ALL DND41 cell line with the expected correlation of shared bound peaks per regulatory element marker (data not shown).

3.3.1.a DBA analysis of DNAse1 Hypersensitivity within the TAL1 locus

Within the *TAL1* locus, the DBA analysis identified sites of differential binding for H3K27ac, DHS and H3K4me3 between different cell samples (Fig 3.4). For DHS, differential binding was identified between immune cells (primary MonoCD14, primary Th1 and Th0, Jurkat and B-cell cell line, CD20RO01794) and other non-immune cell lines (8988T and SAEC) and primary CD4+ T-cells (Th1 and Th0) vs Jurkat (Fig. 3.4, DHS: L Vs O, blue bar; T Vs T, purple bar, respectively). These differential binding sites map to regions 3 and 4, displaying differential binding at intergenic sites upstream of the *TAL1* locus (FDR 0.00403 for I vs O and 0.03 and 0.04 for T vs T, respectively) (Fig 3.4, DHS: 3 and 4 and Table 3.2). This indicates that this DBA is specific to immune cells but may be influenced by the Jurkat cell line as it also shows differential binding at this site relative to primary CD4+ T-cells (Fig 3.4 – DHS, region 3). One other region of DBA was found downstream the intergenic enhancer (*MuTE* Insertion) between T-cell cell lines and T-ALL (Jurkat), indicating a Jurkat cell-specific mark (Fig 3.4 – DHS, region 4).

The mapping of the *TAL1* region was also done using the ENCODE project DHS data for the cell lines, Jurkat, CD20+, CD14+, CD4+ Naïve (T-helper 0), SAEC (normal lung epithelial), 8988T (Pancreas Adenocarcinoma) and Th1 (Fig 3.5). A pattern of open chromatin marker enrichment was found for the *TAL1* region within intragenic regions (not identified by DBA) of the gene and the intergenic *MuTE* insertion region (identified by DBA) (regions 1, 2 and 4), however, are much narrower and more focused within the intragenic region (Fig 3.5 - second red box). In specifically the Jurkat cell line, a large range of read depth of DHS which was found to be statistically

significant from DBA (Fig 3.4 – maps to regions 3 and 4), can be seen relative to other samples tested from the same laboratory (Jurkat, CD20+, CD14+, CD4+ Naïve and SAEC) within the intergenic region highlighted (*), as well as compared to the other cell lines tested (other laboratories) (Fig 3.5 – right box) (see also Appendix for all replicates – Supp. Fig 7.4). However, within the intragenic region of the *TAL1* gene that maps to GeneHancer predicted sites (first and second red box – map to regions 1 and 2 (Fig 3.4)), peaks within the CD14+ (2 samples), CD20+ and 8988T cell lines can also be seen in which DBA found no statistical differences between these cell lines (Fig 3.5, Fig 3.4 – DHS and Table 3.2).

| Table 3.2. DBA Analysis of the *TAL1* locus.

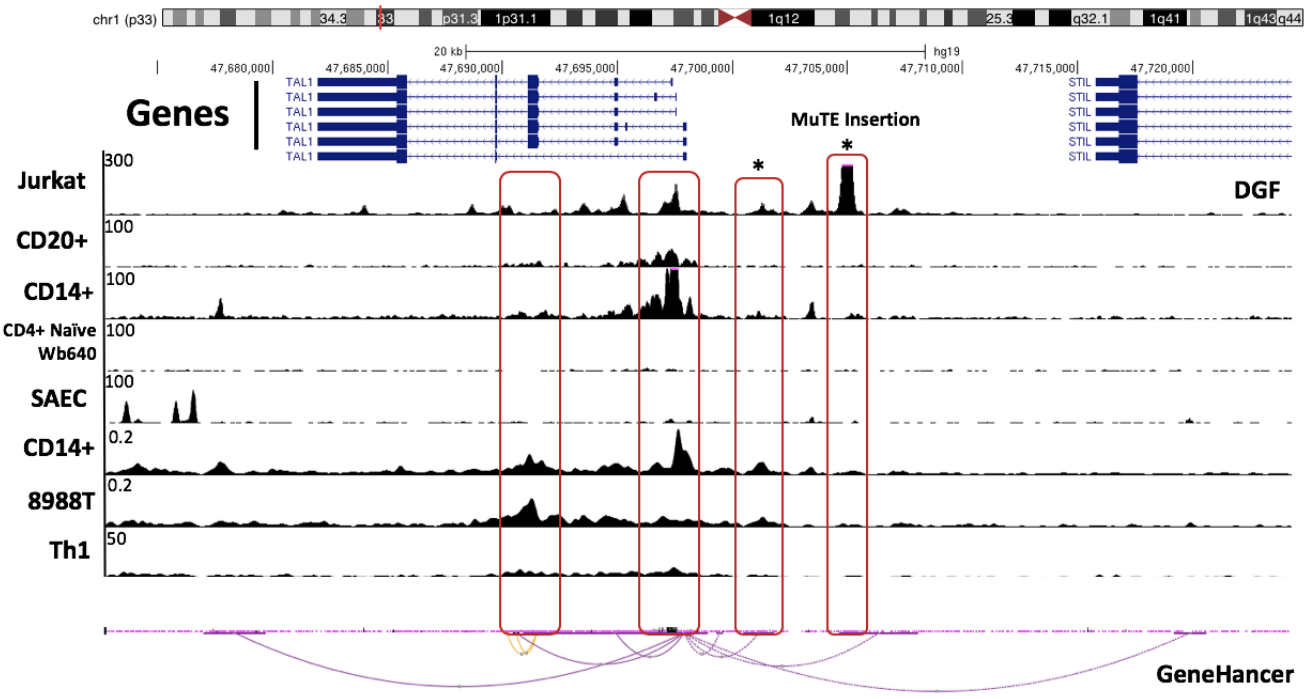
Marker	Region[1]	Contrast[2]	Fold Difference[3]	p-value	FDR
DHS	3	TvsT	1.77	0.00133	0.0286
	4	TvsT	2.89	0.00248	0.0462
	4	IvsO	3.23	0.000369	0.00403
H3K27ac	1	LvsO	3.19	0.0151	0.0293
	2	LvsO	4.37	0.000688	0.00189
	3 and 4	LvsO	4.87	0.00105	0.00275
	1	TvsT	6.39	$6.48E^{-10}$	$2.25E^{-09}$
	2	TvsT	6.26	$3.15E^{-13}$	$1.54E^{-12}$
	3 and 4	TvsT	7.52	$9.86E^{-15}$	$5.52E^{-14}$
H3K4me3	2	LvsO	9.76	0.000101	0.00121
	3	LvsO	5.52	0.00194	0.00978
	4	LvsO	4.64	0.00768	0.0258
	2	TvsT	7.06	$7.54E^{-28}$	$2.34E^{-25}$
	3	TvsT	3.01	0.00452	0.0105
	2	JvsD	10.77	$4.23E^{-25}$	$4.08E^{-23}$
	3	JvsD	6.66	$1.81E^{-08}$	$1.73E^{-07}$
	4	JvsD	5.59	$2.00E^{-05}$	0.00011

[1] Region: refers to the boxed regions in Fig 3.4

[2] Contrast: Contrasts conducted between immune primary cells and cell lines vs non-immune cell lines (I vs O) (DHS) **or** lymphocyte primary cells and cell lines vs non-immune cell lines (LvsO) (H3K27ac and H3K4me3),

primary CD4+ T-cells vs the Jurkat cell line (T vs T) and Jurkat vs DND41 cell line (JvsD) across the *TAL1* locus – specific cell lines listed in Fig 3.4.

[3] Fold difference: The difference between groups contrasted for analysis in which a positive value means a fold increase for immune primary cells and cell lines (IvsO), lymphocyte primary cells and cell lines (LvsO), T-ALL Jurkat (TvsT) and Jurkat (JvsD).



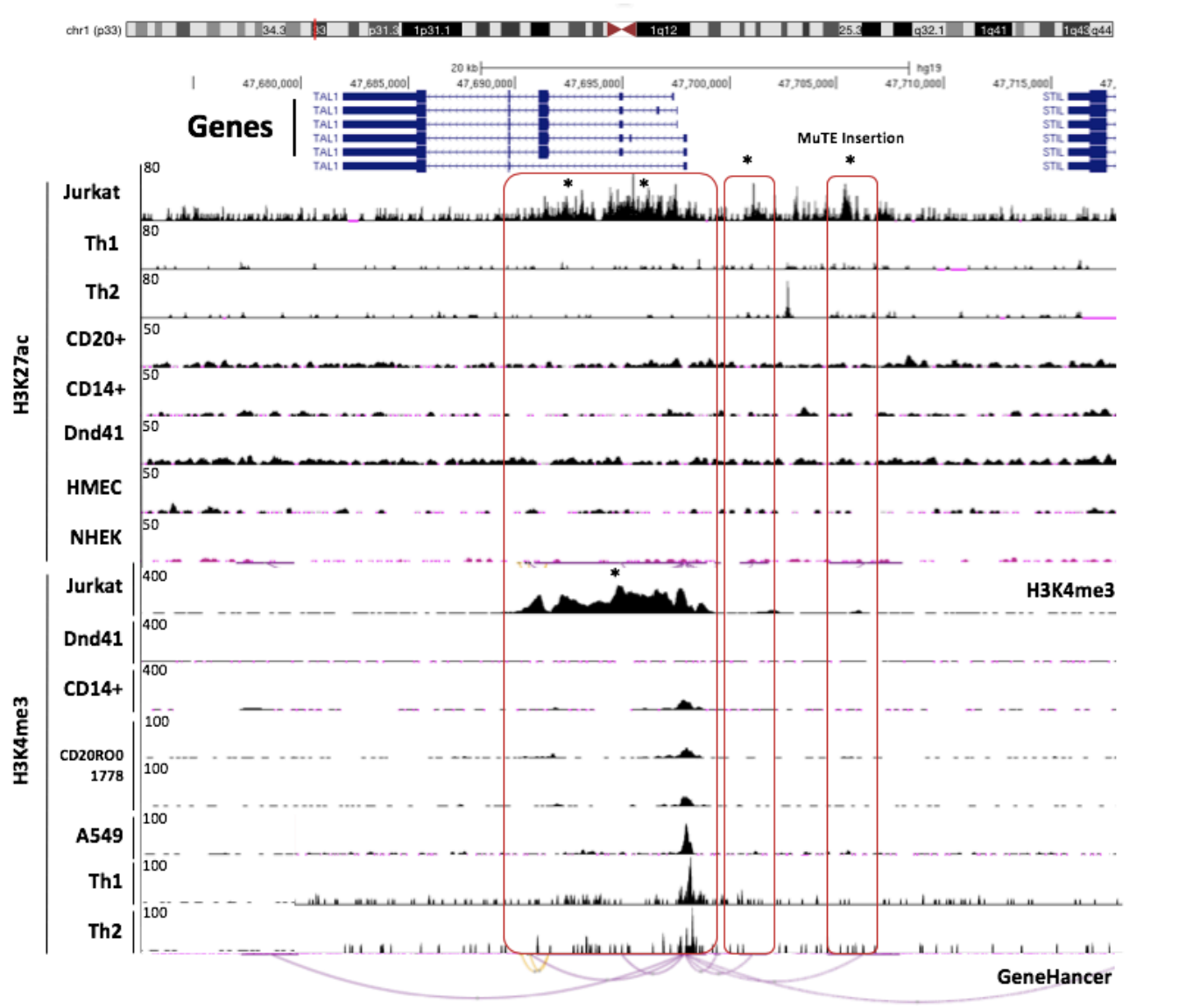
| Figure 3.5. DNAse1 hypersensitivity (DHS) read depth displayed in the UCSC genome browser for the Jurkat, CD20R017794, MonoCD14, CD4⁺ Naïve Wb11970640, SAEC, 8988T and Th1 cell lines across the *TAL1* locus.

Chromosome ideogram of chromosome 1 indicates location of the *TAL1* locus indicates location of the *TAL1* locus indicated by the red line. Ref-Seq genes from the ENCODE Project display isoforms of the *TAL1* and *STIL* gene with localised ChIP-seq peaks of the cell lines listed above. Y-axis displays the read depth of DHS enrichment and is normalised amongst samples from the same experiment (see also Appendix for details and visualisation of all replicates – Appendix Supp. Fig 7.5). GeneHancer database looping displays relationships of regulatory elements across the *TAL1* locus (Methods: 3.2.5). Red boxes highlight areas of DHS enrichment of the Jurkat cell line relative to the other cell lines displayed. DBA analysis has identified regions of statistical difference between cell lines at 2 intergenic regions, one of which is located at the *MuTE* insertion (*).

3.3.1.b DBA analysis of H3K27ac within the TAL1 locus

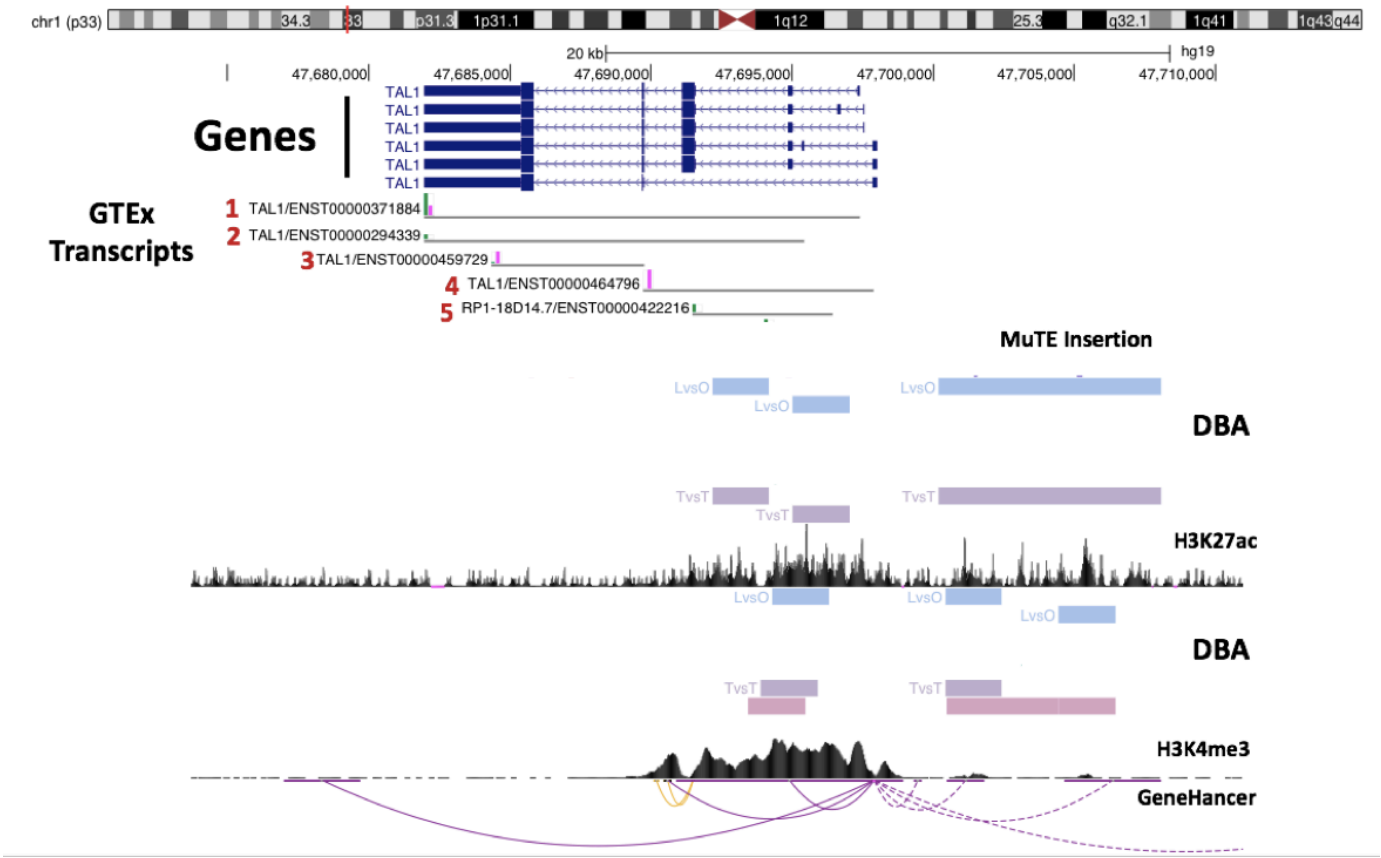
DBA was also conducted for H3K27ac enrichment, which in contrast to the DBA analysis of DHS, identified two intragenic regions (region 1 and 2) in addition to the intergenic region (region 3-4) (Fig 3.4). DBA was conducted between lymphocytes (Jurkat, primary Th1 and 2, and B-cell line CD20RO01794) vs other non-immune cell lines (NHEK and HMEC) (L vs O – blue box) (FDR 0.0293, 0.002 and 0.003 for LvsO, respectively). DBA was also conducted between primary CD4+ T-cells (Th1 and Th2) (T vs T – purple box) (FDR $2.23e^{-09}$, $1.54e^{-12}$ and $5.52e^{14}$, respectively) in contrast to Jurkat within two intragenic sites (Fig 3.4, H3K27ac: region 1 and 2) and the intergenic site (region 3) spanning the *MuTE* insertion enhancer for ~7kbs (Fig 3.4, H3K27ac). This analysis resulted in both DNA contrasts displaying the same regions of differentially bound H3K27ac. Further analysis was conducted between Jurkat and the T-ALL cell line, DND41, for DBA of H3K27ac enrichment within the *TAL1* locus, which identified no statistical differences between these two cell lines. This suggests a similar pattern of enrichment of this intragenic site between the T-ALL cell lines (DND41 and Jurkat) tested.

Mapping of H3K27ac displayed a higher enrichment of Jurkat H3K27ac across the *TAL1* region as identified by DBA, specifically highlighted within gene body regions and the intergenic region upstream of *TAL1* as indicated by the red boxes (Fig 3.6 – H3K27ac, Fig 3.4 – map to regions 1 and 2). This display of H3K27ac can be seen to map to exonic and intronic regions of the *TAL1* gene (over ~15kb), displaying characteristics of open chromatin and regulatory element activity, relative to the other cell lines displayed (Fig 3.6 – H3K27ac). Three regions of H3K27ac enrichment were detected in Jurkat T-ALL samples (also map to Fig 3.4 – regions 1 and 2 and between 3 and 4). The first of these begun at the 5' end of the gene and spread through the first intron and second exon (Fig 3.6, left red box). This region also co-localised with a



| Figure 3.6. H3K27ac and H3K4me3 read depth across the *TAL1* locus.

Chromosome ideogram of chromosome 1 indicates location of the *TAL1* locus indicated by the red line. Ref-Seq genes from the ENCODE Project display isoforms of the *TAL1* and *STIL* gene with localised ChIP-seq peaks of the cell lines listed above. Y-axis displays the read depth of H3K27ac and H3K4me3 enrichment and is normalised amongst samples from the same experiment, with primary cells and cell line samples indicated on the left-hand side (see also Appendix for details and visualisation of all replicates – Supp. Fig 7.4). GeneHancer database looping displays predicted interactions between regulatory elements across the *TAL1* locus (Methods: 2.3.5). Red boxes highlight areas of H3K27ac and H3K4me3 enrichment of the Jurkat cell line relative to the other cell lines displayed, DBA regions are highlighted within these boxed regions (*).



| Figure 3.7. UCSC genome browser display of the *TAL1* locus and mapped GTEx transcripts for thyroid and whole blood cell types (green and pink bars, respectively) with peaks of H3K27ac and H3K4me3 and DBA analyses between cell type groups tested.

Ref-Seq genes display *TAL1* gene isoforms and GTEx transcripts predicted indicate transcription start sites (TSSs) for *TAL1* isoforms. Likelihood of TSS for the tissues, thyroid and whole blood, is indicated by the height of the (green and pink, respectively) where the 5' end of the transcript (grey line) indicates the location of the TSS. Chromosome 1 ideogram displays the location of the *TAL1* locus (indicated by red line).

broad H3K4me3 signal in Jurkat T-ALL samples, and more localised peaks in other cell lines (Fig 3.6 – H3K4me3). Using the GTEx track in the UCSC genome browser possible TSSs can be predicted in the context of multiple isoforms for the *TAL1* gene (https://genome.ucsc.edu/gtex.html, Fig 3.7) and demonstrated transcriptional activity within nearby exons for thyroid and whole blood cell types (Fig 3.7, green and pink bars, respectively). At the 5′ end of the *TAL1* gene, mapping of full isoform transcripts that correspond for whole blood and thyroid specific cell types are seen (bar height indicates transcripts per kilobase million (TPM)) (Fig 3.7 – full gene isoforms 1, 2 and 4 – TPM: 5, 0.5, 0 for thyroid and 1.1, 0, 3.5 for whole blood, respectively), suggesting the 5′ region locates the primary promoter of the gene.

A second region of H3K27ac enrichment mapped upstream of the *TAL1* gene, but did not co-localise with H3K4me3 (Fig 3.6, right red box). This second region also corresponds to the location of the *MuTE* insertion mutation associated with the Jurkat T-ALL super enhancer (Mansour et al., 2015) (Fig 3.6 – H3K27ac). These highlighted regions of enrichment (*) are also seen to have looping relationships to other elements within the region as identified by the GeneHancer looping (Fig 3.6, bottom panel), such as the intragenic region within *TAL1* and the intergenic region of the Jurkat super-enhancer (Fig 3.6 - *). Therefore, a clear relationship between these highlighted regions is seen through looping from the TSS of common *TAL1* transcripts (GTEx transcripts at the 5′ end (Fig 3.7)) to the upstream H3K27ac site (Fig 3.4 - regions 3 and 4).

3.3.1.c DBA analysis of H3K4me3 within the TAL1 locus

For enrichment of H3K4me3, DBA was conducted and identified the 3 regions of DBA (Fig 3.4, H3K4me3 – Regions 2, 3 and 4), specifically displaying differential binding of 2 intergenic sites (including the Jurkat *MuTE* enhancer, regions 2 and 3) and one intragenic region (region 2) between the two sites identified for H3K27ac (Fig 3.4,

H3K4me3). Contrasts between lymphocyte cells and cell lines (primary Th1 and Th2 and Jurkat) vs other non-immune cell lines (MCF7 and SAEC) (FDR 0.00121, 0.0098 and 0.03, respectively), and between Jurkat and DND41 cell lines are found at all three sites (Fig 3.4, H3K4me3 - regions 2-4) (FDR $4.08e^{-23}$, $1.73e^{-07}$ and 0.0011, respectively). However, differential binding is found at the intergenic region downstream of the *MuTE* insertion enhancer (region 3) and within the intragenic region (region 2) for T-cells (Primary Th1 and Th2) and T-ALL (Jurkat) (Fig. 3.4, H3K4me3 - regions 2 and 3) (FDR $2.34e^{-25}$ and 0.01, respectively). No differential binding of H3K4me3 at the Jurkat (*MuTE*) enhancer between primary CD4+ T-cells vs T-ALL (Jurkat) at this site was seen.

The mapping of the promoter marker H3K4me3 was used to confirm the location of the active promoter regions relative to the active enhancer regions identified amongst the Jurkat cell line and the other cell lines tested (Fig 3.6 and Fig 3.4 -H3K4me3). The mapping of H3K4me3 enrichment across the *TAL1* region identified a broad range of H3K4me3 across the intragenic region of the *TAL1* gene (first red box) (Fig 3.6 – H3K4me3) which maps across region 2 for H3K4me3 DBA (Fig 3.4, H3K4me3 – region 2). This region has a broad range of read depth for H3K4me3 across ~10kb which is not seen for the cell lines, DND41, primary cells MonoCD14+, CD20ORO01778 (B-cell), A549 (lung epithelial cancer), primary Th1 and Th2 (Fig 3.6 – H3K4me3). Within this region, a narrow peak of H3K4me3 can be seen for all cell lines (except DND41) which has displayed a putative promoter site (Rivera-Reyes et al., 2016), which also maps to the 5′ end of the GTEx transcripts for TSS location of *TAL1* isoforms for thyroid and whole blood tissue types (Fig 3.7 – full gene transcripts 1, 2, and 4). A contrast can be seen with enhancer markers being restricted to the intergenic region of *TAL1* in this analysis relative to the findings of the *MuTE* super-enhancer stretched across from the intergenic Jurkat enhancer into the first intronic regions of the *TAL1* gene (Mansour et al., 2014).

The enrichment of H3K4me3 in combination with predicted isoform transcripts that are not at this site (upstream), suggest that the DBA found that the intragenic region may be a weak promoter as suggested by the TPM scores of 0.5 and 0 for isoform 3 and 1.1 and 0 for isoform 5 for thyroid and whole blood cell types that map to this site (Fig 3.7 – isoforms 3 and 5). The mapping of H3K4me3 also displays limited H3K4me3 enrichment within the *MuTE* insertion region where the intergenic Jurkat enhancer is located, confirming that this region is a putative enhancer site due to minimal H3K4me3 enrichment relative to the enrichment of H3K27ac and DHS at the same site (Fig 3.4 and 3.6 – right box). DBA between primary T-cells and DND41 compared to Jurkat indicated statistically increased H4K4me3 enrichment (Table 3.2), however in combination with relative H3K4me3 from other cell lines that lack H3K4me3 at this site and relative increases in Jurkat H3K27ac and DHS, this statistical increase is due to regulatory element activity at this site, whereas other cell lines exhibit no signs of regulatory element activity (Fig 3.6 and Fig 3.4 – H3K4me3). Therefore, expected patterns of high H3K27ac and DHS relative to low H3K4me3 indicates region 3 and 4 (Fig 3.4, H3K4me3 and Fig 3.6, H4K3me3 – two right boxes) are putative enhancer sites.

All the patterns displayed amongst the three regulatory elements markers, H3K27ac, DHS and H3K4me3, displayed statistically significant increases in fold change for lymphocyte primary cells and cell lines (which includes Jurkat) and for Jurkat (T vs T and J vs D) for all the stated markers (Fig 3.4 and Table 3.2). This indicates that the Jurkat cell line may have affected the results and influenced the immune or lymphocyte specific signature relative to other non-immune cells lines. Each site mapped had a statistical difference that was lower than a threshold FDR of 0.05, (default from 'DiffBind') with an increased fold difference of 1.77 to 10.77 for immune primary cells/cell lines or lymphocyte primary cells/cell lines (compared to non-immune cell lines), and Jurkat (compared to primary CD4+ T-cells and the DND41 cell

line) (Table 3.2). For H3K27ac and H3K4me3 enrichment, differential binding between T-cell cells and Jurkat as well as Jurkat and DND41 are more significant relative to other regions of DBA found (Table 3.2 - FDR). Overall, the binding profile of regulatory markers, DHS, H3K27ac and H3K4me3 show differential binding within the *TAL1* locus that is specific for the Jurkat cell line.

3.4. Conclusion

Overall, the bioinformatics pipeline and analysis conducted allowed for the mapping of intragenic and intergenic regions within *TAL1*, to display the regulatory landscape through markers of H3K27ac, DHS and H3K4me3. The four regions found were between the first and second exon of the *TAL1* gene (Fig 3.4 - regions 1 and 2) and upstream in the intergenic region, mapping to an intergenic enhancer downstream of the Jurkat *MuTE* super-enhancer site (Fig 3.4 – regions 3 and 4). It was found that these regions have differential binding between lymphocytic groups and other non-immune cell types which was narrowed down between primary T-cells in comparison to the Jurkat cell line and the T-ALL cell line, DND41 and Jurkat. These sites specifically mapped to the putative enhancer region (*MuTE* enhancer) established in literature within the *TAL1* locus, however also displayed a possible Jurkat T-ALL specific region of promoter enrichment within the intragenic region of *TAL1*, in contrast to other studies stating this region is a part of the *MuTE* putative super enhancer (Mansour et al., 2014). This analysis aided in the understanding of the regulatory element landscape of the *TAL1* locus and can be used to map other epigenetic markers, such as DNA methylation, to help target testing of differential DNA methylation at regulatory sites between Jurkat clonal populations established in Chapter 2.

Chapter 4 - DNA methylation and genetic variation across the *TAL1* locus

4.1. Introduction

4.1.1. DNA methylation

DNA methylation and histone subunit methylation and acetylation modulate patterns of gene expression across the genome without the alteration of a DNA sequence (Moore et al., 2013). DNA methylation involves the transfer of a methyl group onto the second cytosine base of a CpG dinucleotide, which are concentrated in clusters within CpG islands (Kulis and Esteller, 2010). CpG islands are regions of DNA that have a higher density of CpG dinucleotide sequences than the rest of the genome, however they are typically unmethylated (Bird et al., 1995). These regions are evolutionarily conserved within humans, promoting the regulation of gene expression through association with chromatin structure and TF binding, specifically at promoters and enhancers (Moore et al., 2013).

At promoters, DNA methylation leads to the silencing of the targeted gene but DNA methylation at intragenic and intergenic enhancers can have varying effects on gene expression (Sharifi-Zarchi et al., 2017). It is believed that intragenic DNA methylation inhibits cryptic or intragenic enhancer-driven transcription enabling efficient transcription from the unmethylated promoter (Sharifi-Zarchi et al., 2017). Loss of intragenic enhancer methylation increases eRNA transcription, which may cause interference with elongating RNA Pol II activity from the promoter of the gene in which it is located, and can also enable gene transcription at neighbouring gene(s)

(Cinghu et al., 2017). Therefore, DNA methylation within the gene body is complex and dynamic.

4.1.1.a. Methylation-Sensitive Restriction Endonuclease (MSRE) assay

A MSRE assay allows for the rapid detection of DNA methylation within multiple amplicons simultaneously (Melnikov et al., 2005). It is a simple method that relies on the digestion of genomic DNA (gDNA) by methylation sensitive restriction endonucleases (Pandey et al., 2016). Although bisulphite-based methods are considered the benchmark in the analysis of DNA methylation, these methods have several limitations that are due to the technically demanding parameters that need to be met and the cost of sequencing technologies (Pandey et al., 2016). However, MSRE assays using qPCR can be used for rapid and simultaneous detection of DNA methylation that allows for high-level multiplexing, minimal assay optimisation, and a low concentration of gDNA needed for each assay (Pandey et al., 2016).

Through the use of commercially available restriction endonucleases, the MSRE assay can be used in a cost-effective way. In the absence of DNA methylation, methylation-sensitive restriction endonucleases will cleave the DNA and the subsequent qPCR amplification will detect high cycle threshold values (CT) due to decreased amplification from the limited starting product (Hashimoto et al., 2007). If a CpG dinucleotide site is methylated, the methylation-sensitive restriction endonuclease cannot cleave the DNA template, resulting in a high yield during the qPCR, with lower CT values (Hashimoto et al., 2007). The digestion of a DNA template by a methylation-sensitive restriction endonuclease is compared to digestion by an isoschizomer that digests both methylated and unmethylated DNA with equal efficiency (Hashimoto et al., 2007). The methylation-sensitive endonuclease HpaII and its isoschizomer MspI both cut at the sequence, CCGG, with methylation at the CpG dinucleotide inhibiting cleavage by HpaII but not MspI. The difference in CT values

of amplicons from gDNA digested with HpaII or MspI can be quantified and used as an assessment of methylation status at that site (Hashimoto et al., 2007).

The aim of this investigation is to use a quantitative MSRE assay to assess the dynamics of DNA methylation at the *TAL1* locus within regulatory sites identified in Chapter 3 (Fig 3.4) in the different Jurkat clonal cell lines. Reduced representation bisulphite sequencing (RRBS) and Illumina HumanMethylation450 array (Methyl450) data from the ENCODE project and the Cancer Cell Line Encyclopaedia (CCLE) will be used to define expected patterns of DNA methylation within the *TAL1* locus and will be compared to results obtained using the MSRE assay.

4.1.2. Nanopore sequencing

Oxford Nanopore Technologies' MinION sequencer is being utilised as a compact and powerful device for long-read sequencing, particularly for long PCR amplicons (thousands to ten thousands of base pairs) (Leggett and Clark, 2017; Orsini et al., 2018). Nanopore sequencing is done at a single-molecule level by detecting an electric current of single-stranded DNA as it passes through a protein channel (nanopore) (Wei et al., 2018). Benefits from using nanopore technology involve sequencing that is focused towards long reads that achieve highly contiguous assembly and identifying structural variation within samples (Bowden et al., 2019). The accessibility of the MinION Nanopore system, and its ability to read long sequences, made it an ideal choice for sequencing PCR-generated amplicons of 2kb or longer that covered much of the *TAL1* locus. Through the generation of libraries of *TAL1* amplicons from the Jurkat parental and clonal cell lines, it was possible to test whether differences seen within the phenotypic, transcriptional and epigenetic landscape of *TAL1* in these cell lines was correlated with genetic variation.

The aims of this chapter are to assess the dynamics of DNA methylation at the *TAL1* locus within regulatory sites identified within Chapter 3 and use the MinION nanopore sequencer for *TAL1* regions to test whether genetic differences exist between the Jurkat parental and clonal cell lines (established in Chapter 2).

4.2. Methods

4.2.1. Bioinformatic analysis of the DNA methylation landscape across *TAL1*

To identify sites of differential DNA methylation within the *TAL1* region, data supplied from the ENCODE project was used to identify targets for the MSRE assay. These data sets are from the HAIB RRBS and HAIB Methyl450 array data conducted by the Myers Lab within the HudsonAlpha Institute (Varley et al., 2013). This data was displayed together with the location of putative regulatory elements identified in Chapter 3 (Chapter 3 - Fig 3.4). RRBS is an approach for large-scale high-resolution DNA methylation analysis through bisulphite conversion and sequencing of MspI digested small CG rich sequences to represent the whole genome (Meissner et al., 2005). The Methyl450 bead array data provides high-throughput DNA methylation quantification through bisulphite conversion and whole-genome amplification across >450, 000 methylation sites (Dedeurwaerder et al., 2014). The FANTOM5 track was also used to map mRNA transcriptional activity from Cap Analysis of Gene Expression sequencing (CAGE-seq) (Noguchi et al., 2017).

Based on the intersection of sites of putative regulatory elements and the sites of DNA methylation, targets were identified for MSRE analysis (Table 4.1). Primers were designed to target these sites by using the UCSC genome browser to retrieve the DNA sequence from seven broad regions of interest. The MspI/HpaII recognition sequence 'CCGG' were identified within each of these seven broad regions. Amplicons between 150-250bp which contained a CCGG sequence located in the middle of the sequence

were selected and primers designed using the Primer 3 program (http://bioinfo.ut.ee/primer3-0.4.0/), ensuring the CCGG sequences were not within the primer sequences. Nineteen primer sets were identified and verified within the UCSC genome browser using the 'In-silico PCR' tool to validate the region of interest to be tested for the MSRE analysis (https://genome.ucsc.edu/cgi-bin/hgPcr). Five of the 19 primer sets were utilised to target exonic, the two Jurkat intragenic sites (Chapter 3 - Fig 3.4), the intergenic H3K27ac enriched site downstream of the *MuTe* insertion and the Jurkat enhancer region (*MuTE* Insertion) of the *TAL1* gene (Table 4.1).

| Table 4.1. List of MSRE primers designed to target CpG dinucleotide sites within the *TAL1* locus.

Primer Name	Coordinates	Forward Primer (5'-3')	Reverse Primer (5'-3')	Target Region
MSRE_1	Chr1:47,685,696-47,685,857	AGGCGGAGGATCTCATTCTT	CTAATCTCCAGGTCCCCACA	Exonic
MSRE_2	Chr1:47,694,782-47,694,979	CTGTCCTGAGCCTTCCTCAC	AAAAAGGGGGAAAGCAAAGA	Downstream TSS Peaks, H3K27ac and H3K4me3 enrichment
MSRE_3	Chr1:47,695,312-47,695,481	GAACATTTTCGAACCCTCCA	CTATTCGCCTTTCCCAACAC	Co-localised TSS Peaks, H3K27ac and H3K4me3 peak
MSRE_4	Chr1:47,701,431-47,701,582	GGTTCTCCCTAAACCCCAAA	ATAAACTCGGCTGCTCATCA	Downstream intergenic H3K27ac and DHS Peak
MSRE_5	Chr1:47,705,095-47,705,268	CGCATGTGCATTCTCTCTGT	TGCCTTGCTTCTATGGGGTA	Jurkat (*MuTe*) enhancer

[1]Primer amplicon sequences in Appendix – 7.3.2.a

DNA methylation within CpG Islands of the *TAL1* locus was mapped using data supplied from the Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al., 2019). *TAL1*-specific regions were retrieved from the CCLE database for T-ALL-specific and leukaemia cell lines out of the 1457 cell lines available (https://portals.broadinstitute.org/ccle/data). The data file

Cell_lines_annotations_20181226.txt was downloaded to identify T-ALL and leukaemia cell lines studied by the CCLE and the data file CCLE_RRBS_cgi_CpG_clusters_20181119.txt was used to retrieve percentages of methylation at the two clusters of CpG islands identified within the *TAL1* gene (chr1:47,690,440 – 47,691,404 and chr1: 47,690,503 – 47,693,585) (Fig 4.5). These regions were displayed in the UCSC genome browser by uploading a BED file that included the coordinates of the two clusters of CpG islands within *TAL1*. The cell line data that was used included: Jurkat (T-ALL), PEER (T-ALL), HL60 (acute myeloid leukaemia), K562 (chronic myelogenous leukaemia), MOLM-13 (acute myeloid leukaemia), ALLSIL (T-ALL), KASUMI-2 (B-ALL), DND41 (T-ALL), HPBALL (T-ALL), PF382 (T-ALL), RPMI8402 (T-ALL), P12-ICHIKAWA (T-ALL), RCHACV (B-ALL) and MOLT-3 (T-ALL). Data was plotted in the web tool 'Methylation plotter' according to the formatting rules described by Mallona et al. (2014). This provided a methylation profile heatmap, a summary of the categorised groups based on cell type, a dendrogram to cluster identified groups, and statistical testing using a Kruskal-Wallis test to assess differences between the groups (http://maplab.imppc.org/methylation_plotter/). This analysis provided an understanding of DNA methylation patterns within intragenic CpG islands in multiple leukaemia cell lines within the *TAL1* gene.

4.2.2. gDNA isolation of Jurkat Clonal Populations

Jurkat clonal populations were grown under the same conditions presented in Chapter 2 (0.1% FBS starvation in 1% PSF RPMI 1640 media for 24 hours and 5% FBS 1% PSF in RPMI 1640 media for 72 hours within 96 round-bottom well plates). Clonal populations P0, C11 and C4 were used for the MSRE assay based on their differential expression of the *TAL1* gene (high, moderate and low, respectively) at passages 1 and 9 (Table 4.2). Passages 1 and 9 were used as they displayed significant differences between clonal populations in the phenotyping and gene expression analysis

conducted in Chapter 2. The cell line, A549 (lung epithelial cancer), was also used to compare differences of methylation status as predicted by UCSC RBBS and Methyl450 data between Jurkat at the exonic and intragenic H3K27ac and H3K4me3 peak region within the *TAL1* gene (Primers MSRE_1 and MSRE_3).

The cell line populations were harvested and gDNA isolation was conducted per the Bioline ISOLATE II Genomic DNA kit protocol (https://www.bioline.com/us/downloads/dl/file/id/873/isolate_ii_genomic_dna_kit_product_manual.pdf). The concentration of isolated gDNA was checked using the Qubit 2.0 Fluorometer using the Qubit dsDNA BR assay kit (ThermoFisher) as stated by the Qubit 2.0 Fluorometer protocol.

| Table 4.2. Averaged duplicates for *TAL1* gene expression (Chapter 2) for Jurkat clonal and parental cell lines at passage 1 and 9 and arranged based on relative *TAL1* expression[2].

Passage 1		Passage 9	
Cell Line	TAL1	Cell Line	TAL1
P0[1]	1.428	P0	1.772
C9	0.882	P00	1.518
C3	0.789	C10	1.282
C2	0.750	C5	0.969
C11	0.748	C3	0.891
P00	0.700	C9	0.666
C1	0.628	C1	0.593
C8	0.600	C11	0.579
C10	0.396	C2	0.567
C4	0.351	C6	0.502
C6	0.222	C8	0.501
C5	0.127	C4 (9)	0.399

[1]Boxed parental and clonal cell lines were utilised for MSRE assay due to consistent *TAL1* expression ranging from high (P0), moderate (C11) and low (C4) between passage 1 and 9 relative to other clonal cell lines.

[2]Heatmap displays red as high expression, white as moderate expression and blue as low expression of *TAL1* (relative fold change).

4.2.3. MSRE Protocol

Commercial restriction endonucleases that are used for MSRE assays include methylation-sensitive (HpaII) and methylation-insensitive (MspI) (Tran et al., 2010). These endonucleases determine methylation at CpG dinucleotides as they cleave at CCGG recognition sites (Tran et al., 2010). The use of both a methylation-sensitive and methylation-insensitive endonucleases allows for the comparison of a reference control sample (methylation-insensitive endonuclease) that will cleave at methylation sites regardless of methylation status, therefore limiting false positive results due to polymorphic variation at restriction sites and incomplete digestion by endonucleases (Tran et al., 2010).

Duplicate digestions for HpaII, MspI and mock digestions were done for gDNA samples (P0, C11 and C4 at passage 1 and 9), with triplicate qPCR assays done for each duplicate digestion (n=6 per cell line and passage) (Table 4.3).

| Table 4.3. Parameters used for gDNA digestion with endonucleases HpaII, MspI and mock digested gDNA.

gDNA Concentration (ng)	150ng (10ng/µL)
Enzyme Concentration (Units)	5U – MspI and HpaII
Reaction Volume (µL)	15µL
Reaction Length (Hours)	1
Heat Inactivation	95°C - 15 minutes

Digestions were conducted in the following reaction: 1 x CutSmart Buffer (New England BioLabs), 1 µL of 5U/µL HpaII and MspI (New England BioLabs), 150ng of gDNA, in a total volume of 15µL. Digestions were incubated at 37°C for 1 hour followed by a 95°C heat inactivation for 15 minutes using the Applied Biosystems Veriti Thermal Cycler. The digested gDNA was stored at -20°C until required for use.

MSRE qPCR analysis of the digested gDNA was done using 10ng of gDNA, 1 x SensiFAST SYBR Green No-ROX Kit from a 2x stock (Bioline), forward and reverse primers at a final concentration of 0.4µM in a total final volume of 10µL. The reaction was run using the Applied Biosystems 7500 Fast Real-Time PCR system with cycling conditions: 95°C for 5 minutes, followed by 40 cycles of 95°C for 10 seconds, 62°C for 10 seconds and 72°C for 30 seconds.

4.2.4. MSRE assay analysis

The MSRE assay was analysed by comparing the dCT values generated from gDNA digested with HpaII or MspI (positive control) and the mock digested DNA (negative control), i.e. dCT HpaII (HpaII - Mock Digestion) and dCT MspI (MspI - Mock Digestion). This was done per digest and primer set and averaged for the triplicate values per digest (n=6 per duplicate digests). The data distribution of dCT HpaII and dCT MspI for each primer set was tested for statistical significance using a two-tailed Wilcoxon Signed-Rank test (https://www.socscistatistics.com/tests/signedranks/default2.aspx) that is a statistical test for non-parametric data for differences of paired values of unequal variance (Nahm, 2016). Statistically significant differences ($p<0.05$) between dCT MspI and dCT HpaII were followed-up for further analysis using box plot outliers.

To assign methylation status to each amplified region, a boxplot analysis of outliers was conducted. This was done according to the method described in Chapter 2.2.6. Lower inner and outer fences for dCT MspI digested gDNA per region were calculated per region and used as thresholds to determine the distribution of unmethylated values (dCT MspI). If dCT HpaII values fell above the lower inner fence this was deemed unmethylated, values between inner and outer fences (mild outlier)

were deemed partial/low DNA methylation, and high methylation was seen when values fell lower than the lower outer fence (extreme outlier).

Statistical correlations of the MSRE dCT HpaII data was also tested using the Spearman's Rho correlation test (https://www.socscistatistics.com/tests/spearman/default.aspx) for correlations between passages 1 and 9 per region tested, per cell line population and for CFSE and *TAL1* gene expression results from Chapter 2.

4.2.5. Amplicon generation

Due to the advantage of nanopore sequencing conducting long-reads, a targeted whole-gene approach was undertaken to generate amplicons using PCR (Orsini et al., 2018). Due to the error-rate of nanopore technologies, PCR amplification of targeted regions can ensure high-level read depth at targeted sites and compensate for this limitation (Orsini et al., 2018). Primers were designed to target chr1:47,691,795-47,705,959 across the *TAL1* locus that mapped to regulatory sites and DNA methylation sites identified in Chapter 3 and 4.

The UCSC genome browser was used to identify regions for MinION Nanopore Sequencing by mapping regions within the *TAL1* locus that span sequences targeted by the MSRE assay amplicons. Using the UCSC genome browser, DNA sequences for these regions were extracted using the 'View' and 'Get DNA' options (https://genome.ucsc.edu/) (Appendix – section 7.3.2.b). The DNA sequence for each region was copied into the Primer3 database (http://bioinfo.ut.ee/primer3-0.4.0/) with the following parameters: Temperature minimum: 59°C, Temperature Optimal: 60°C, Temperature maximum: 61°C for a product size of 1500 – 4000bp. Primers were selected to generate overlapping amplicons (~200bp) and amplicons tested and

selected using the UCSC genome browser 'In-Silico PCR' option (https://genome.ucsc.edu/cgi-bin/hgPcr) (Table 4.4).

|Table 4.4. Primers designed for MinION Nanopore sequencing for targeted regions across the *TAL1* locus.

Primer Name	Coordinates	Forward Primer (5'-3')	Reverse Primer (5'-3')
Seq_1	Chr1:47,691,795-47,694,488	CAGAGACTGAGGGCCAAAAG	CTCGAGTGGGCCCTGAT
Seq_2	Chr1:47,694,309-47,697,743	ACCCAGCCTCTGGTCTCTCT	AGAGGTCTTCGCTCCCTTTC
Seq_3	Chr1:47,697,580-47,701,354	CGCGCATTTCTGTATATTGC	CCCAGTTCCAGACTCCAGAC
Seq_4	Chr1: 47,699,942 – 47,702,280	AGGAAACAGCTGGACCAAGG	AGTAGTAGGCCTGGGGTGAG
Seq_5	Chr1:47,701,548-47,704,837	GCCATGCATGCACTCTGATG	AGGGCTCCAGGGTATGCTAA
Seq_6	Chr1:47,703,476-47,705,959	GGCTCATCTCACCCAGTCAC	CTGCCTCTCCTTCTCACTGC

[1]Amplicon sequence information in Appendix – section 7.3.2.b

The primers were tested for the appropriate annealing temperature using the New England BioLabs Tm calculator (https://tmcalculator.neb.com/#!/main) for the Q5 High-Fidelity DNA polymerase (New England BioLabs) at a final primer concentration of 0.5µM. All annealing temperatures were noted and used to conduct the generation of *TAL1* amplicons.

Amplicons were generated using the Q5 High-Fidelity DNA polymerase (New England BioLabs) protocol (https://www.nebiolabs.com.au/protocols/2013/12/13/pcr-using-q5-high-fidelity-dna-polymerase-m0491). gDNA was isolated as described (Methods: 4.2.2). All reactions were assembled on ice and quickly transferred to a pre-heated Applied Biosystems Veriti Thermocycler at 98°C. A touch-down PCR was used to increase the specificity and sensitivity of the PCR amplification. The touch-down PCR technique increases the initial annealing temperature above that of the predicted primer annealing temperature and decreases the annealing temperature over multiple cycles (Korbie and Mattick, 2008).

In some cases, successful PCR required the addition of 10% dimethylsulfoxide (DMSO). The following reaction was set up dependent on the amplicon: 5x Q5 Reaction buffer (New England BioLabs) at a 1x concentration, 10mM dNTPs at a final concentration of 0.2mM (Bioline), forward and reverse primers at a final concentration of 0.5µM each (Sigma-Aldrich), 75ng of gDNA, 0.5µL of Q5 High-Fidelity Polymerase (20 units/mL) (New England BioLabs), DMSO (as required, 10%) in a final volume of 50µL. The PCR amplification was then run as described in Table 4.5.

| Table 4.5. PCR cycling conditions for *TAL1* amplicons using the Q5 High-Fidelity polymerase.

Step	Temperature (°C)	Time
Initial Denaturation	98	2 minutes
10 cycles	98	10 seconds
	(T_{AP} + 10°C) or (T_{AP} + 10°C) -5°C 1°C decrease per cycle	10 seconds
	72	40 seconds/kb
25 cycles	98	10 seconds
	T_{AP} or T_{AP} -5°C	10 seconds
	72	40 sec/kb
Final Extension	72	2 minutes
Hold	10	-

*T_{AP}: predicted annealing temperature as calculated by https://tmcalculator.neb.com/#!/main.

Amplification of amplicons was verified using a 1.2% agarose gel stained with SYBR Safe DNA gel stain (ThermoFisher Scientific) in 1x Tris-borate-EDTA (TBE) buffer within an electrophoresis tank (BioRad). Samples were loaded with 5x DNA loading buffer (Bioline) at a 1x concentration and run with the Hyperladder 1kb (Bioline).

4.2.6. Nanopore library preparation and sequencing

After confirmation of the amplification of targeted *TAL1* regions, library preparation was conducted. First, concentrations of gDNA from the amplicons generated was measured using the Qubit 2.0 Fluorometer and Qubit 2.0 BR assay kit (ThermoFisher Scientific) (as per manufacturers' instructions – Chapter 2). Amplicons were combined into cell line-specific pools at a concentration of 0.2 pM for each amplicon and made up to 100µL with 1x Tris-EDTA (TE) buffer (10mM Tris-HCl, pH 8.0, 1mM EDTA). The gDNA concentration for each amplicon pool was measured using the Qubit 2.0 fluorometer and then purified using the Monarch DNA Cleanup and Gel Extraction Kit (New England BioLabs), eluted in 40µL and measured for gDNA concentration again (pools were between 0.6-1.2µg in total). Next, to prepare the amplicons for ligation of the sequencing barcodes, end repair of the amplicons was conducted using the NEBNext Ultra II End Repair/dA-Tailing Module (New England BioLabs). The reagents were mixed in the following order: 3µL of NEBNext Ultra II End Prep Enzyme mix, 7µL of NEBNext Ultra II End Prep Reaction buffer, 50µL of DNA (made up with TE for a of maximum gDNA amount of 1µg) for a total volume of 60µL. Reaction was run in an Applied Biosystems Veriti Thermocycler with the cycling conditions: 20°C for 30 minutes, 65°C for 30 minutes and held at 4°C.

After end repair, a column purification was conducted (Monarch DNA Cleanup and Gel extraction kit) before ligation of barcode adaptors. Using the Blunt/TA ligase Master Mix (New England BioLabs) and the Nanopore PCR Barcoding Expansion Kit (Oxford Nanopore Technologies), ligation was conducted at a ratio of 5:1 (adaptors : fragments) with 12µL of purified End Prep DNA (300ng), 8µL of barcode adaptors and 20µL of Blunt/TA Ligase Master Mix (in this order). Reaction was mixed gently, and contents of the tube were spun down. Bead purification was done using 60µL of resuspended Agencourt AMPure XP DNA purification beads (Beckham Coulter) and mixed thoroughly. Reactions were incubated on a rotator mixer for 5 minutes at room

temperature, spun down and pelleted using the DynaMag-2 Magnetic Particle Concentrator (Invitrogen). Supernatant was discarded after pelleting and beads were washed using 200µL of 70% ethanol (Sigma-Aldrich) (freshly made) and without disturbing the pellet, ethanol was removed and the washing with ethanol was repeated. The sample tube was briefly centrifuged and placed on the magnet and residual ethanol was evaporated and left to dry at 37°C with lid open. Pellet was resuspended in 25µL of ddH_2O, incubated for 2 minutes at room temperature, and placed on a magnet to pellet beads (eluate to be clear and colourless). Then 15µL of eluate was retrieved and quantified using the Qubit 2.0 fluorometer. The library was then diluted to 10ng/µL with ddH_20 or TE Buffer.

Barcoded DNA was amplified as follows: 5x Long Amp Reaction buffer (New England BioLabs), final concentration of 0.3mM dNTPS (New England BioLabs), PCR barcode primers at a final concentration of 0.4µM (BC01-BC06) (Oxford Nanopore Technologies), 2µL of Long Amp Taq polymerase (5U in total) (New England BioLabs), 20ng adapter ligated template in a final volume of 50µL. The PCR reaction was run with the following cycling conditions: Initial denaturation: for 3 minutes at 95°C, denaturation for 15 seconds at 95°C and annealing for 15 seconds at 62°C and extension for 50 seconds/kb at 65°C (15 cycles), extension for 10 minutes at 65°C and held at 4°C. Barcoded amplicons were quantified using the Qubit 2.0 Fluorometer and run on a 1.2% agarose E-Gel with SYBR Safe DNA gel stain to confirm amplification of pooled amplicons. Barcoded templates were pooled in equimolar ratios (in most cases, a maximum amount of approximately 400ng of gDNA per primer set, with each library containing 6 primers) and purified using the Agencourt AMPure XP DNA purification beads (as stated above). An end repair step was not done as the Long Amp polymerase adds an overhanging A base (New England Biolabs). Next, adaptor ligation was conducted using 60µL of gDNA sample from previous step, 25µL of thawed Ligation Buffer (Oxford Nanopore Technologies), 10µL of NEBNext Quick T4

DNA Ligase (New England BioLabs) and 5µL of Adaptor Mix (Oxford Nanopore Technologies) and incubated for 10 minutes at room temperature. We next conducted the AMPure XP bead purification, however used 250µL of Short Fragment Buffer (Oxford Nanopore Technologies) to wash the beads. Beads were pelleted again, supernatant was discarded and repeated and allowed to dry for 30 seconds. Next, 15µL of Elution Buffer was used (Oxford Nanopore Technologies) and incubated for 10 minutes at room temperature, beads were pelleted, and the eluate was retained in a clean 1.5mL centrifuge tube. Ligated DNA was assessed for concentration using the Qubit 2.0 Fluorometer. DNA library was loaded at a concentration of 50 fmol.

After library preparation, priming of the MinION flow cell was done using the Flow Cell Priming Kit (Oxford Nanopore Technologies) and conducted as per manufacturers' instructions. The MinION sequencer was run for 16 hours before termination of the run and extraction of the data using the MinIT companion (Oxford Nanopore Technologies) that conducts base-calling of the data through the Guppy algorithm and outputs FAST5 and FASTQ files for analysis.

4.2.7. Nanopore Sequencing Analysis

FASTQ files generated from MinION Nanopore sequencing of Jurkat populations P0, C4 and C11 across the *TAL1* locus required processing before conducting genetic variant analysis. Using the Python package, 'qcat' (https://github.com/nanoporetech/qcat), FASTQ files were combined into a single FASTQ file, trimmed and demultiplexed to remove adaptors and sorted into the barcoded samples (P0, C11 and C4 at passage 1 and 9) (EPI2ME algorithm for demultiplexing (ONT))(Deshpande et al., 2019). Next, the Python package 'Minimap2' was used for alignment of long reads of DNA (specific for nanopore sequencing) against a reference genome (Li, 2018). This resulted in an aligned SAM file to be

converted to a BAM file for analysis. Using 'Samtools' within Python, SAM files were converted to BAM files and sorted by coordinate (as conducted in Chapter 3 Methods – 3.2.3). BAM files processed and sorted were then tested for quality metrics and read length distributions using the European Galaxy database option 'Nanoplot' (Theuns et al., 2018). All command lines can be found in Appendix – 7.4.3.

Also within the 'Samtools' package, the 'Bcftools' component can be used to compute the genotype likelihood of samples, call variants and provide statistics of variants in a binary variant call format (BCF) (Li, 2011; Li et al., 2009). 'Bcftools' also operates faster than other variant calling programs and is able to generate multi-way pileups, producing genotype likelihoods (mpileup). BCF files were converted to VCF files and analysed for sequence variants (Danecek and McCarthy, 2017). This identified genetic variants that differed from the reference genome (Human Genome Assembly GRCh37/hg19) with metrics of genotyping, allele depth (unfiltered number of reads per allele), depth of coverage (filtered depth of reads per allele), SP (Phred-scaled strand bias p-value) and allelic depth per strand (ADF (Forward) and ADR (Reverse)) within a VCF File format as described by the Broad Institute (https://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it). VCF files were analysed per sample to identify single-nucleotide variants (SNVs) within the Jurkat cell lines and were narrowed down based on a threshold of < 60 for Phred-scaled strand bias p-value (SP). SP was used as it ensures that genetic variants identified have an equal amount of read depth on both strands to ensure correct evaluation of the alleles counted per strand as described by the Broad Institute (https://software.broadinstitute.org/gatk/documentation/article?id=6925). The identified SNVs were verified visually using the Integrative Genome Viewer (IGV). Sorted BAM files were loaded onto IGV, as was the VCF file that contained the genotyping information from the mpileup from 'Bcftools'. This allowed for a direct

comparison of 'Bcftools' genotyping calls with aligned Nanopore sequencing data. These sites were also mapped with the 1000 Genomes Project Phase 3 single-nucleotide polymorphisms (SNPs) (from the IGV server) to identify common and rare variants (The 1000 Genomes Project Consortium, 2015; Zheng-Bradley and Flicek, 2017). The Genome Aggregation Database (gnomAD) by the Broad Institute was also used to retrieve known variants for the *TAL1* locus (https://gnomad.broadinstitute.org/) (Karczewski et al., 2019) (Table 4.9). A normalised BIGWIG track for DHS within the Jurkat cell line (ENCODE data) that identifies open chromatin at regulatory DNA sequences was also included to assist in localising variants that reside within the regulatory landscape of *TAL1*. All command lines can be found in Appendix – 7.4.3.

We then tested the linkage disequilibrium (LD) of identified SNPs that co-localise Jurkat SNVs using the NCI LDLink (https://ldlink.nci.nih.gov/) database. Using the 'LDProxy' option, genome-wide association studies (GWAS) SNPs were entered and downloaded as a .csv file to identify if SNPs identified to map Jurkat SNVs are proxy SNPs with a LD correlation of $r^2 > 0.8$ (Carlson et al., 2004; Teo et al., 2009). Proxy SNPs in LD were analysed for Utah Residents from North and West Europe (CEU) population groups only. The r^2 measure was used as it represents allelic frequency of pairwise LD SNPs more accurately than D' values which are also used for LD analysis (VanLiere and Rosenberg, 2008).

A set of criteria was placed to predict SNV functional relevance within regulatory elements. SNVs were displayed in the UCSC genome browser to show co-localisation of predicted TF binding sites (TFBS) from the JASPAR 2020 database track as validation for possible functional applications of co-localising Jurkat SNVs with TF binding (Khan et al., 2018b), as well as the Genomic Evolutionary Rate Profiling (GERP) to map highly conserved regions of the genome. This identified two SNVs that

may have functional relevance for regulatory element dysfunction within all Jurkat cell lines and were further tested through predictive analysis of 810 TF binding profiles within the JASPAR 2020 data base (http://jaspar.genereg.net/). The reference allele and alternative allele for the two SNVs were tested and displayed differential binding profiles at the SNV site and 12-bp upstream and downstream to mimic TF binding at sites from 6-12bp (Tuğrul et al., 2015).

4.3. Results

4.3.1. MSRE Optimisation

To ensure the MSRE assay was reproducible and quantitative, the following conditions were optimised: gDNA concentration, HpaII and MspI endonuclease concentration, reaction volume and reaction time for digestion with MspI and HpaII and the qPCR cycling parameters were also optimised for this assay.

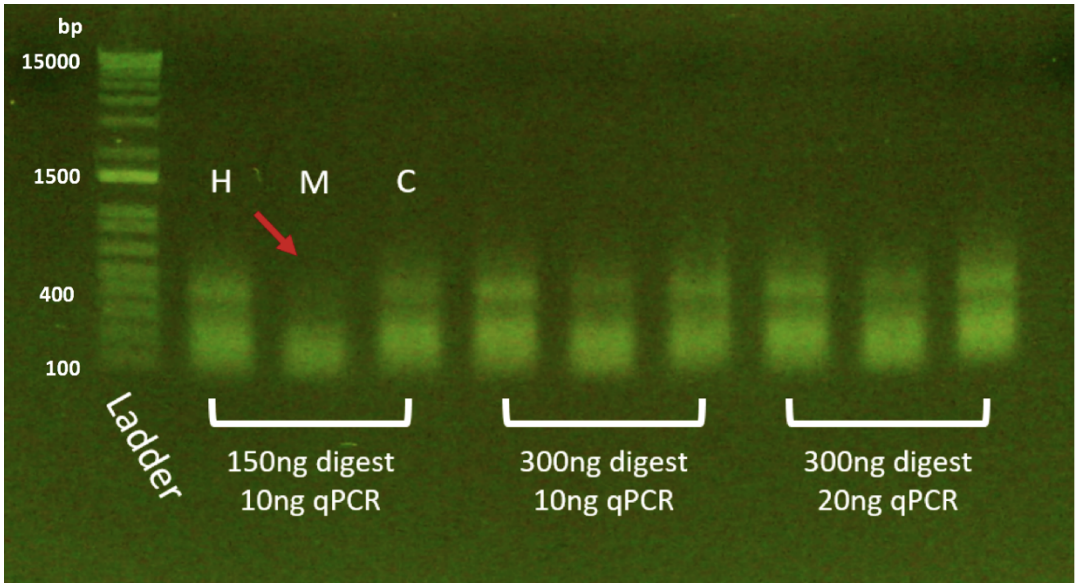
Initially, the following digestion conditions were tested based on protocols from literature (Table 4.6) (Hashimoto et al., 2007; Holemon et al., 2007; Oakes et al., 2006).

| Table 4.6. Digestion optimisation conditions with endonucleases HpaII and MspI for Jurkat cell line gDNA.

MSRE Optimising Conditions		
gDNA Concentration (ng)	300 (20ng/µL)	150ng (10ng/µl)
Enzyme Concentration (Units)	5	
Reaction Volume (µL)	15	
Reaction Length (Hours)[1]	1	
Other[2]	15 min 95°C Heat Inactivation	

[1] New England Biolabs recommend a one hour digest for a gDNA concentration of 1µg (https://www.nebiolabs.com.au/protocols/2012/12/07/optimizing-restriction-endonuclease-reactions).

[2] New England BioLabs suggest a heat inactivation step to inhibit STAR activity of endonucleases such as HpaII, which is inactivated at 80°C, whereas MspI cannot be heat inactivated.



| Figure 4.1 Digested gDNA with endonucleases HpaII, MspI and mock digested at digestion concentrations of 150ng and 300ng and qPCR gDNA concentration with 10ng or 20ng.

Red arrow indicates a low yield of amplicon product at 150ng of digested gDNA and 10ng of digested gDNA in the qPCR reaction. E-Gel 1Kb Plus DNA ladder (ThermoFisher Scientific) was used to measure amplicon size from 100 – 15000bp.

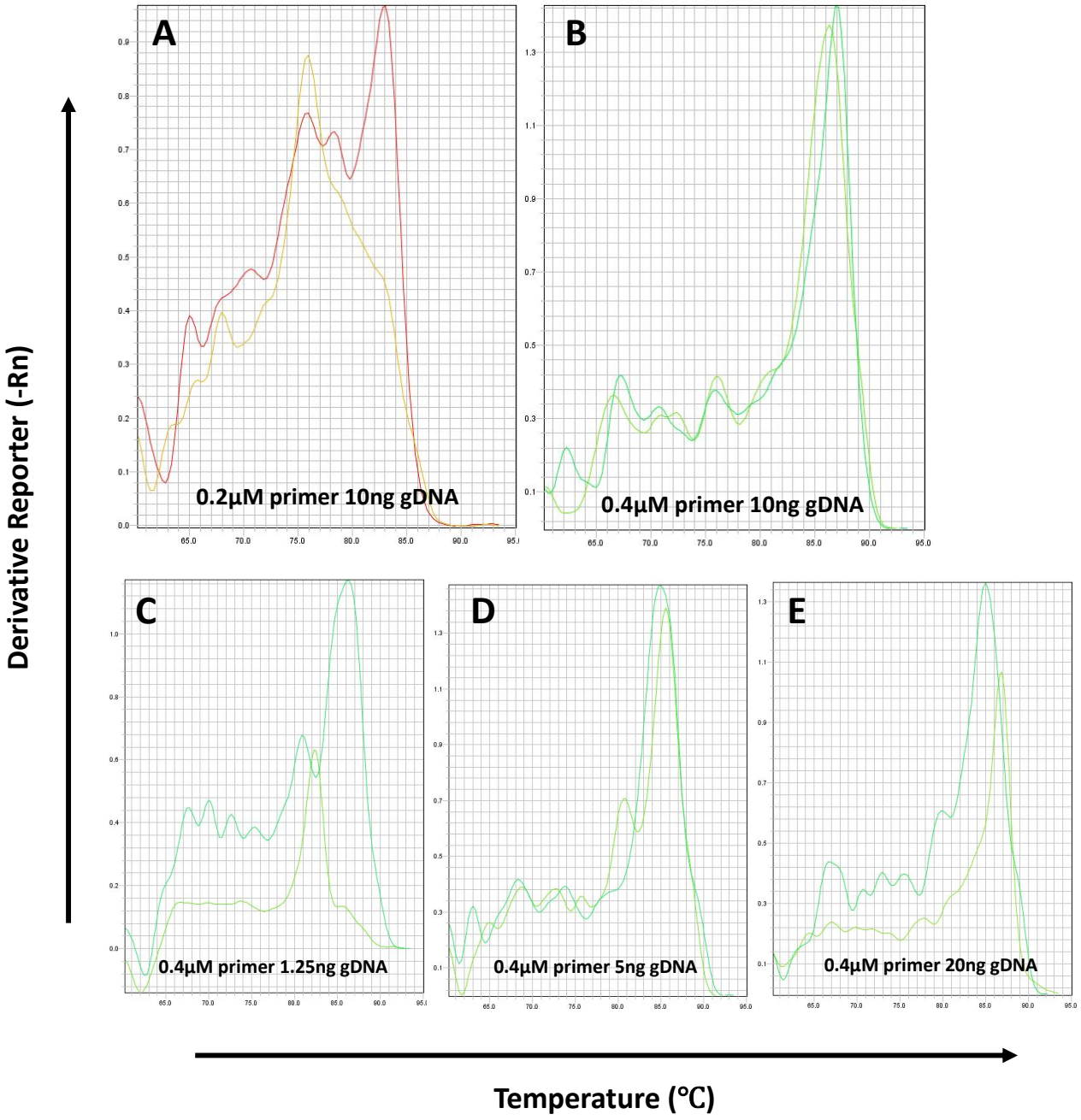
The very low amplicon yield associated with a starting digestion of 150ng of gDNA and a total of 10ng in the qPCR reaction (Fig 4.1, red arrow) indicated the optimal condition for digestion of gDNA by MspI. However, despite defining optimal digestion of gDNA with MspI, qPCR CT values displayed a large variation amongst all HpaII, MspI and Mock digested samples (Standard Deviation > 1) (data not shown).

To improve the sensitivity of the assay we next tested different concentrations of primers (0.2 and 0.4μM final concentrations) and used a 3-step qPCR method (Table 4.7) that is reported to be more accurate when compared to the standard 2-step PCR (Fast qPCR) (Hilscher, 2005).

| Table 4.7. Previous 2-step qPCR cycling conditions changed to the 3-step qPCR cycling conditions for a more sensitive MSRE assay.

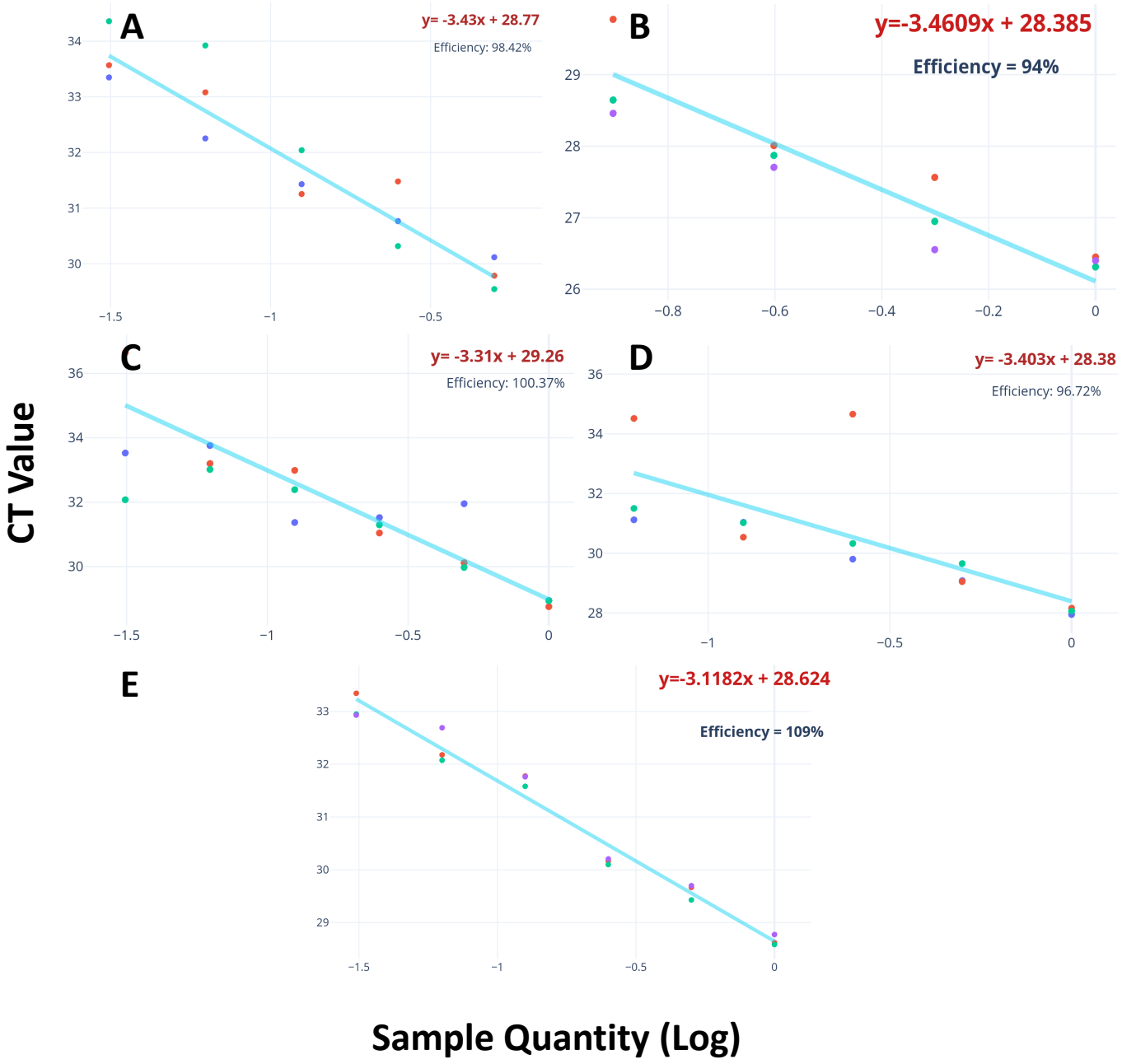
Previous 2-step qPCR Cycling Conditions			
Cycle	**Temperature (°C)**	**Time**	**Step**
1	95	2 min	Polymerase activation
2-40	95	3 sec	Denaturation
	60	30 sec	Annealing and Extension
Changed to 3-step qPCR cycling Conditions			
1	95	5 min	Polymerase Activation
2-40	95	10 sec	Denaturation
	62	10 sec	Annealing
	72	30 sec	Extension

The 3-Step qPCR method was tested with a starting amount of 150ng of digested gDNA with 1.25, 2.5, 5, 10 and 20ng of gDNA in a qPCR reaction and 0.2μM or 0.4μM final primer concentrations. Differences between 0.2μM and 0.4μM were distinguished by the melt curve analysis which that displayed a single melt peak at



| Figure 4.2. (A-B) Melt Curve plot of duplicate amplicons generated through MSRE optimisation.

Optimisation for MSRE primer 2 (Table 4.1) 10ng of MspI digested gDNA (150ng) within the qPCR reaction at a final primer concentration of 0.4µM and 0.2µM, respectively. (C-E). Melt curve plot of duplicate amplicons generate through MSRE optimisation for MSRE primer 2 at 1.25, 5 and 20ng of MspI digested gDNA (150ng) within the qPCR reaction with a final primer concentration of 0.4µM (respectively). Derivative reporter (negative first derivative of normalised fluorescence (Rn) by the reporter) on the x-axis and temperature on the y-axis. A single peak indicates the 'melting' of a single amplicon product after qPCR MSRE analysis.



| Figure 4.3. (A-E) MSRE Primer Efficiencies.

Scatter plots of triplicate CT values (y-axis) against log sample quantity (Dilution factor 2) (x-axis) for MSRE primers, MSRE_1, MSRE_2, MSRE_3, MSRE_4 and MSRE_5, respectively (10ng of mock digested gDNA with a final primer concentration of 0.4µM in reaction). The slope equation of the standard curve (light blue line) is displayed (red) which was used to test the primer efficiency (black) per primer set (Methods 4.2.3.b.).

0.4μM, whereas multiple melt peaks are seen at 0.2μM (Fig 4.2A,B) (Melt curve cycling as stated in Chapter 2 Methods: 2.2.10). It is clear that multiple peaks appear at lower concentrations of gDNA using a final concentration of 0.4μM primers (Fig 4.2C and D) and that variability between replicates at 20ng of gDNA was also displayed (Fig 4.2E). Therefore, the following MSRE assay was conducted with 10ng of digested gDNA using a final concentration of 0.4μM for the primers (Fig 4.2B).

Having established gDNA digestion and amplification conditions, we next tested the amplification efficiency for each primer set (Table 4.1). Primers at a final concentration of 0.4μM per reaction were tested with mock digested gDNA with 1:1 serial dilutions of 2.5 to 0.156 ng, confirming each primer pair was amplified at between 90-110% efficiency (Fig. 4.3) (Rogers-Broadway and Karteris, 2015).

4.3.2. The DNA methylation landscape of *TAL1*

The prediction of DNA methylation sites across the *TAL1* locus was done through the use of ENCODE supplied data for RRBS and Methyl450 bead array data (Methods: 4.2.1). Methylation data was mapped across the *TAL1* locus along with markers for active regulatory DNA: H3K27ac, DHS and H3K4me3 (Fig 4.4). The combined analysis of DNA methylation and histone methylation and acetylation identified five regions for MSRE analysis (Fig 4.4, pink vertical lines, MSRE_1-5). These regions targeted the 3' exon (Fig 4.4, MSRE_1), two intragenic sites (Fig. 4.4, MSRE_2 and 3), an intergenic site immediately downstream of the *MuTE* insertion (Fig 4.4, MSRE_4) and the Jurkat *MuTE* insertion and intergenic enhancer (Fig. 4.4, MSRE_5).

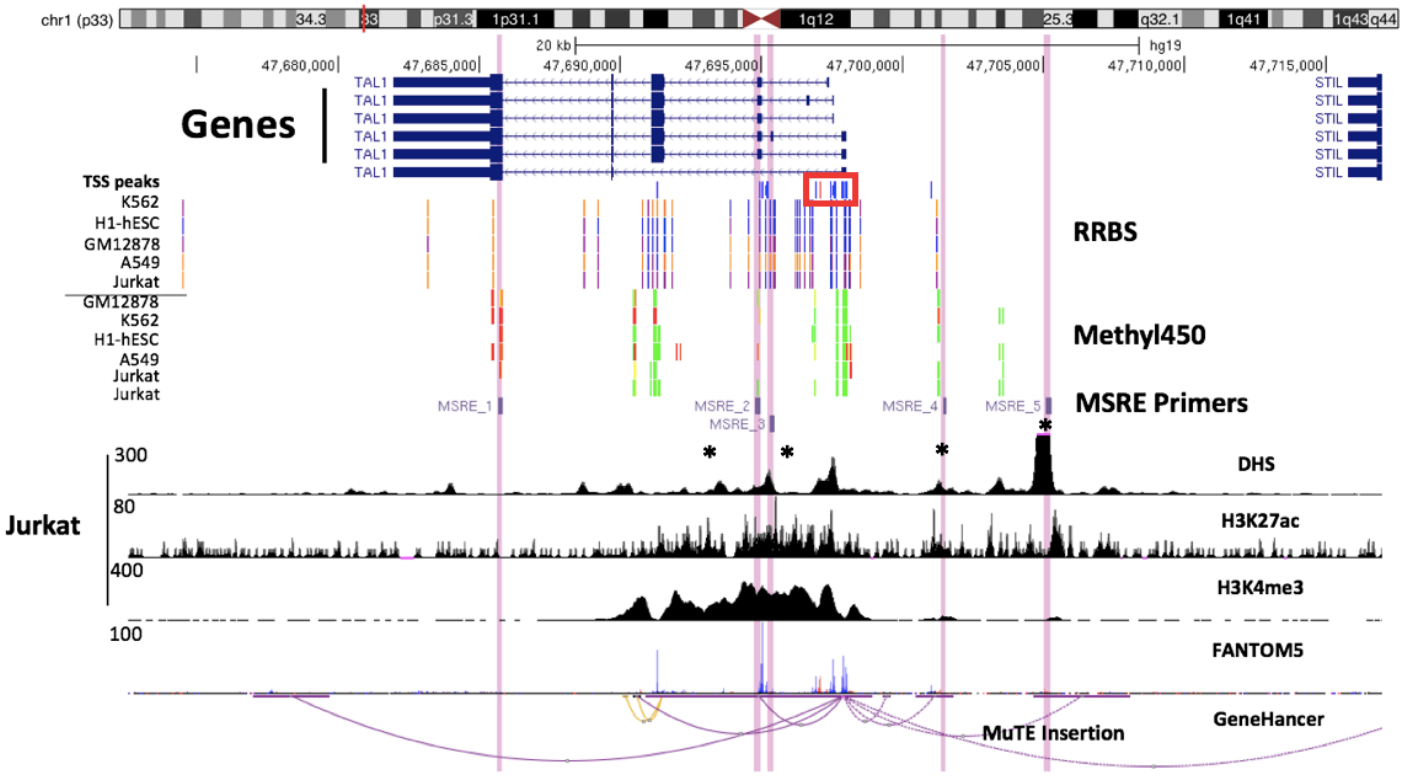
In Chapter 3, sites of intragenic and intergenic histone acetylation and methylation and the relative transcripts from the predicted TSS were utilised to identify DNA methylation for MSRE analysis. Specifically, MSRE_2-5 match DBA peaks of DHS, H3K27ac and H3K4me3 (Fig 4.4, Jurkat - *). Within the two intragenic sites (Fig 4.4,

MSRE_2 and 3) that are downstream of the exon, GTEx transcripts are displayed for full isoforms of the *TAL1* gene within thyroid and whole blood cell types (Fig 4.4, TSS peaks, red box). H3K27ac and DHS DBA peaks at the two intergenic sites that map enriched H3K27ac and DHS downstream of the *MuTE* insertion (Fig 4.4, MSRE_4) and the Jurkat *MuTE* insertion and intergenic enhancer (Fig 4.4, MSRE_5) were also used for the MSRE analysis.

Using the RRBS and Methyl450 data, patterns of partial or low methylation (blue/purple and green bars) of the *TAL1* locus in Jurkat T-ALL cells was seen in comparison to the A549 cell line (methylated) (Fig 4.4 – RRBS and Methyl450, respectively), except at the 3' exon (Fig 4.4, MSRE_1) which is methylated for both Jurkat and A549 (Fig. 4.4, RRBS, Methyl450, Jurkat and A549 – red/orange bars). Therefore, the combination of regulatory element markers using DBA from Chapter 3 in conjunction with RRBS and Methyl450 data has allowed for the targeting of MSRE sites 1 to 5 (Fig 4.4, MSRE Primers).

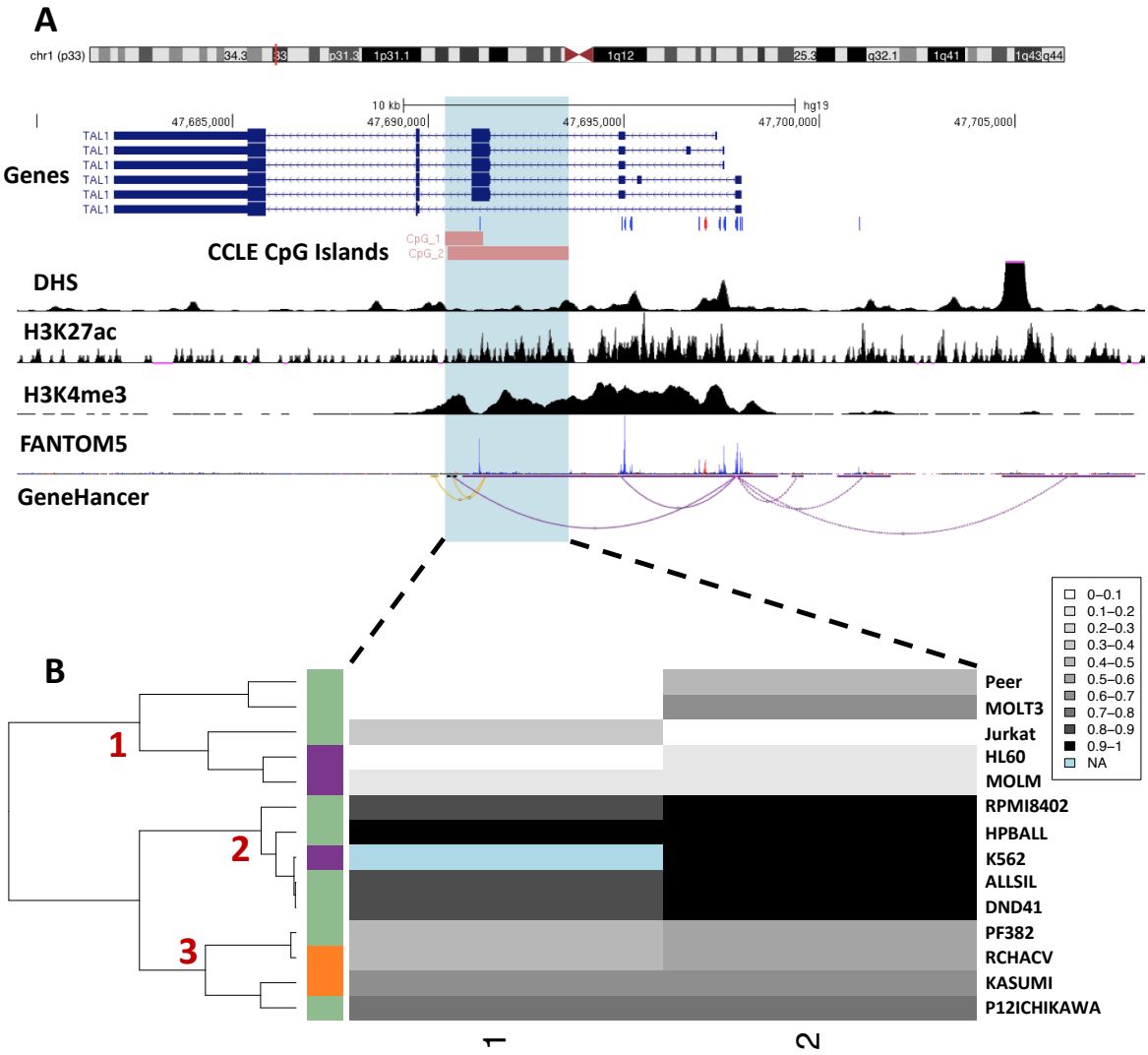
Using RRBS data from the CCLE (Ghandi et al., 2019), DNA methylation across the CpG island in *TAL1* was also assessed (Fig 4.5A). Using data from 14 different cell lines, comprised of T-ALL (Fig. 4.5B, Green bars - Peer, MOLT3, Jurkat, RPMI6402, HPBALL, ALLSIL, DND41, PF382, P12ICHIKAWA), myeloid leukaemia (Fig. 4.5B, purple bars - HL60, MOLM, K562) and B-ALL (Fig 4.5B, orange bars - RCHAV, KASUMI), the percent methylation of each cell line was plotted using hierarchical clustering (Fig 4.5B). Relative to the other cell lines tested, the T-ALL Jurkat cell line was hypomethylated at the CpG islands and was similar to Peer, MOLT3, HL60 and MOLM cell lines (Fig 4.5B), and coincides with literature describing hypomethylation within CpG sites within *TAL1* in late cortical T-ALL (Haider et al., 2018) (Fig 4.5B – cluster 1, green bars) (Haider et al., 2018). However, other T-ALL cell lines such as RPMI8402, HPBALL, DND41 and ALLSIL displayed a hypermethylation pattern

across both CpG islands (Fig 4.5B, cluster 2, 0.8-1). The myeloid leukaemia cell lines and B-ALL cell lines tend to group together within their respective groupings and



| Figure 4.4. Identification of MSRE Target Regions.

Highlighted MSRE primer targets localised with methylation sites identified by HAIB RRBS and HAIB Methyl450 bead array data supplied by ENCODE through the UCSC genome browser. HAIB RRBS data displays DNA Methylation of cell lines, K562, H1-hESC, GM12878, A549 and Jurkat. RRBS DNA methylation displays unmethylated (blue), partial methylation (purple) and methylation (orange). HAIB Methyl450 displays cell lines, GM12878, K562, H1-hESC, A549 and 2 replicates of the Jurkat cell line. Methyl450 displays unmethylated (green), partial methylation (yellow) and methylation (red). These DNA methylation markers are mapped to Jurkat regulatory elements identified in Chapter 3 using DBA (H3K27ac, H3K4me3 and DHS), as well as the Jurkat enhancer (*MuTE* Insertion) (*). Chromosome 1 ideogram displays location of the *TAL1* region (red line) and Ref-Seq genes display isoforms of the genes, *TAL1* and *STIL*. GeneHancer looping supplied from ENCODE is also shown with FANTOM5 mRNA transcript sites (red: forward strand, blue: reverse strand), red box indicates common transcript sites from GTEx data of thyroid and whole blood cell types



| Figure 4.5. Display of CCLE tested CpG Island sites within the *TAL1* gene and percent methylation.

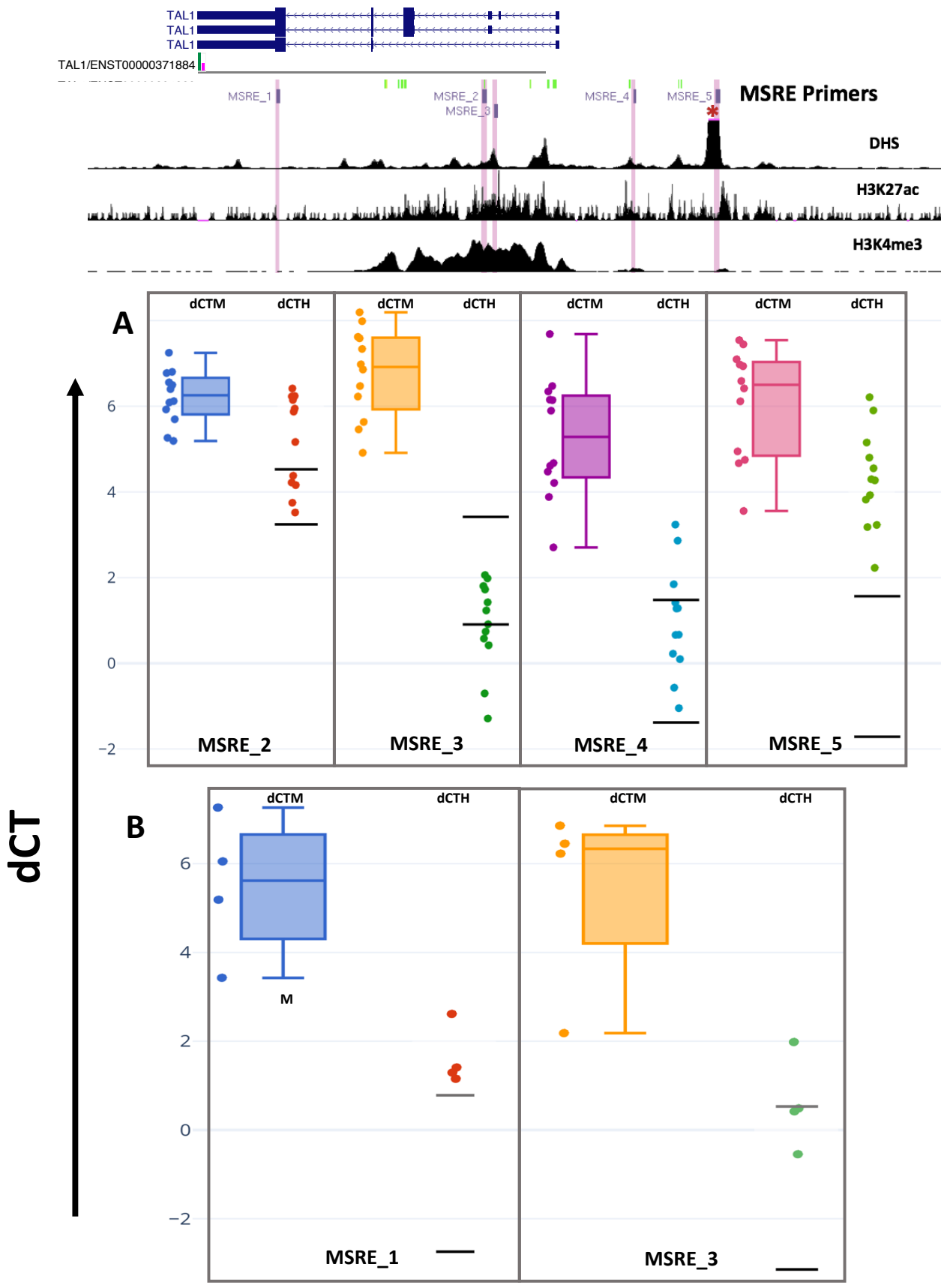
A.*TAL1* regions (CpG 1 and 2) with localisation of Ref-Seq genes for *TAL1*, regulatory ChIP-seq (H3K27ac and H3K4me3) and DHS markers, FANTOM5 mRNA transcript sites (red: forward strand, blue: reverse strand) and GeneHancer looping within the UCSC genome browser. **B.** Hierarchical clustering heatmap of CCLE RRBS data for DNA methylation at *TAL1* CpG islands 1 and 2 (left and right, respectively) for cell lines Peer, MOLT3, Jurkat, HL60, MOLM, RPMI6402, HPBALL, K562, ALLSIL, DND41, PF382, RCHAV, KASUMI and P12ICHIKAWA. Clustering dendrogram displays the grouping of T-ALL (green), myeloid leukaemia (purple) and B-ALL (orange) cell lines based on DNA methylation percentage. Clustering dendrogram identifies methylation trends of tested cell lines into 3 clusters, cluster 1: hypomethylation/partial (0-0.5), cluster 2: partial/hypermethylation (0.8-1) and cluster 3: partial methylation (0.3-0.6). Kruskal-Wallis statistical testing displays no significant difference between groups ($p < 0.05$) (see also Appendix – Supp. Table 7.27). Scale of DNA methylation percentage is indicated in the heatmap legend (top left).

have shown hypomethylation or moderate methylation across these CpG sites, respectively (Fig 4.5B – purple and orange groups). Although no statistically significant differences were found between the different cell line groups for DNA methylation at the *TAL1* CpG islands (Appendix - Supp. Table 7.27), it was apparent that T-ALL methylation is heterogeneous (Fig 4.5B), and that in the context of Jurkat-specific T-ALL, intragenic CpG Islands within *TAL1* are hypomethylated.

4.3.3. MSRE assay on Jurkat parental and clonal cell lines (P0, C11 and C4) for regions across the *TAL1* locus

With this understanding of the predicted DNA methylation landscape across *TAL1*, an MSRE assay was done using DNA extracted from the Jurkat parental (P0) and clonal cell lines C11 and C4 as these cell lines exhibited high, moderate and low levels of *TAL1* expression (Table 4.2). These Jurkat cell lines were tested at MSRE sites 2-5, which represented two intragenic (Fig 4.4, MSRE_2 and 3) as well as two 5' intergenic sites (Fig 4.4, MSRE_4 and 5). In addition, validation of the MSRE assay against published methylation data sets was done using Jurkat and A549 cell lines (Fig. 4.4, MSRE_1 and 3).

Using a boxplot outlier analysis, the dCT for MspI – Mock (dCT MspI) and HpaII – Mock (dCT HpaII) were displayed as averages of each replicate digest (Fig 4.6A and B - P0, C11, C4, passage numbers 1 and 9, n=12). Since dCT MspI values represented digested (i.e. unmethylated) MSRE amplicons, and the mock digested control represented undigested amplicons (i.e. methylated), the values for dCT MspI (CT MspI – CT mock) must be well above zero. Based on this rationale, an MSRE site was identified as methylated if the dCT HpaII values (CT HpaII – CT mock) clustered around zero. The assigning of methylated/not methylated to the MSRE amplicons was formalised by using outlier analysis of a box plot: lower inner and outer fences were calculated for each MSRE dCT MspI dataset. The dCT HpaII values were plotted and, relative to the MspI data, were defined as mild outliers when mapped between the


TAL1
TAL1
TAL1
TAL1/ENST00000371884
MSRE_1
MSRE_2
MSRE_3
MSRE_4
MSRE_5
MSRE Primers
DHS
H3K27ac
H3K4me3
A
dCTM
dCTH
6
4
2
0
−2
MSRE_2
MSRE_3
MSRE_4
MSRE_5
dCT
B
M
MSRE_1
MSRE_3

| Figure 4.6. MSRE Analysis of the *TAL1* locus in Jurkat Cell Lines.

(Previous Page). Location of MSRE primers 1-5 (MSRE_1-5) relative to H3K27ac, DHS, H3K4me3 and the longest isoform transcript identified by GTEx for thyroid and whole blood cell types (green and pink bars, respectively. **(A)** Methylation status of the intragenic (MSRE_2 and 3) and 5' extragenic (MSRE_4 and 5) using Jurkat cell lines P0, C11 and C4. Boxplots show results of dCT MspI (dCTM) with results of dCT HpaII shown as scattered data points (dCTH) (**B**) Validation of the MSRE assay using Jurkat (P0) and A549 gDNA samples (MSRE_1 and 3). Boxplot inner and outer lower fences calculated from the dCT MspI datasets (n=12 in each case) are shown as black lines: unmethylated data falls above lower inner fence (top line), partial methylated samples between the inner and outer fences, and high methylation samples below the lower outer fence. X-axis displays region tested corresponding to the targeted regions highlight*ed* (top panel) and y-axis displays HpaII dCT values.

inner and outer fences; and defined as extreme outliers when mapped below the lower outer fence. In this way, MSRE amplicons could be identified as not methylated (above inner fence), low/partial methylation (values between inner and outer fences) and high methylation (below outer fence).

The box plot analysis displayed the distribution of data of the positive control i.e. the expected CT value for an unmethylated samples (dCT MspI) relative to the dCT values of each sample per region (dCT HpaII).

We first analysed the two intragenic MSRE sites (Fig 4.4, MSRE_2 and 3), which map to intragenic TSS sites (GTEx transcript) and peaks of H3K4me3 and H3K27ac (Fig 4.4). dCT HpaII values are mainly distributed between the unmethylated and partial methylated states (Fig 4.6A, MSRE_2: CT MspI boxplots shown in blue, CT HpaII values show as red points; inner and outer lower fences indicated as black lines). This region is downstream of DBA identified H3K27ac and H3K4me3 intragenic enrichment based on Chapter 3 DBA (Chapter 3 - Fig 3.6) and GTEx transcripts from thyroid and whole blood cell types identified that major *TAL1* isoform transcripts (full

transcripts) are upstream of this site (Chapter 3 - Fig 3.7). Inspection of the HpaII data points showed the parental and clonal cell line 11 shifted from unmethylated at passage 1 to partial methylation at passage 9 whereas the opposite was seen for cell line C4 (Table 4.8).

However, this is different for the second intragenic site that maps H3K27ac and H3K4me3 and TSS transcripts (MSRE_3) that co-localise within an active regulatory element peak. In this case, the distribution dCT HpaII values (Fig. 4.6, MSRE_3, dCT HpaII) falls within the partially methylated and high methylation status (Fig4.6A, MSRE_3: dCT MspI shown in yellow, CT HpaII values show as green data points; inner and outer lower fences indicated as black lines). Therefore, inspection of the dCT HpaII data points displayed the parental cell line shifting from a partially methylated state to a high methylation state at passage 1 (Table 4.8 – P0). The opposite is seen for clonal cell line 4 where the methylation state changes from high methylation to partially methylated between passage 1 and 9, whereas clonal cell line 11 remains partially methylated between passages (Table 4.8 – C11 and C4). Therefore, despite the relative proximity of the MSRE_2 and MSRE_3 regions to each other (Fig 4.4), the changes in DNA methylation between MSRE_2 and MSRE_3 were evident, shifting from an overall unmethylated state to a partial/methylated state. Thus, a dynamic shift of DNA methylation occurs at these sites, particularly at MSRE_3, and these sites are localised to intragenic peaks of H4K3me3 and H3K27ac.

Next, we analysed the methylation status of the two intergenic MSRE sites, which include the *MuTE* insertion enhancer (Fig 4.4, MSRE_5), and a downstream site that is defined by increased H3K27ac and DHS in Jurkat T-cells (Fig 4.4, MSRE_4). At the intergenic H3K27ac and DHS enriched region downstream of the *MuTE* insertion (MSRE_4), dCT HpaII values mainly fell within the partially methylated and unmethylated states (Fig 4.6A, MSRE_4: dCT MspI shown in purple, dCT HpaII

values shown as light blue data points; inner and outer lower fences indicated as black lines). Prior DBA of this site indicated the presence of a Jurkat specific H3K27ac and DHS enriched region that does not co-localise with the *MuTE* insertion (Chapter 3 – Fig 3.4). At this site, partial methylation was seen amongst all clonal cell lines, except clonal cell line 11 at passage 1 (Table 4.8). This indicated an overall stable state of DNA methylation across cell lines and passages at this intergenic H3K27ac and DHS enriched site.

In contrast, the *MuTE* insertion enhancer (MSRE_5) displayed dCT values that fell within the unmethylated state for all cell lines at both passages (Fig 4.6A, MSRE_5: dCT MspI shown in pink, dCT HpaII values are shown as light green data points; inner and outer lower fences indicated as black lines). This site localises with the *MuTE* insertion mutation that introduces MYB binding motifs and harbours TAL1 CRC binding activity (Mansour et al., 2014). In addition, DBA identified the differential binding at this region as statistically significant for the Jurkat cell line, relative to other cell lines tested in Chapter 3 (Fig 3.4). The stable unmethylated state seen between all cell lines and passages (passage 1 and 9 correlation tested using Spearman's Rho test, $p < 0.05$) suggested that this enhancer was active in all Jurkat cell lines, in comparison to the putative downstream intergenic enhancer (Fig 4.6 - MSRE_5 Vs MSRE_4). This result also validated the MSRE assay by identifying low DNA methylation that agreed with the function of this super-enhancer site within T-ALL.

In order to compare the DNA methylation status of the *TAL1* locus presented here to published methylation data from cell lines in other laboratories, comparisons of differential methylation were tested using the A549 cell line (lung epithelial cancer) and the Jurkat parental cell line (P0) at the exonic and intragenic H3K27ac and H3K4me3 peak site (MSRE_1 and 3) to HAIB RRBS and HAIB Methyl450 data (Fig 4.6B). This showed different results to those seen within the ENCODE HAIB RRBS

and Methyl450 data for the exonic site and intragenic peak site for A549 and Jurkat (Fig 4.5 and 4.6B – MSRE_1 and 3). In our MSRE assay, both the A549 and the Jurkat cell line DNA was unmethylated at the exonic site (Fig 4.6B - MSRE_1: dCT MspI shown in blue, dCT HpaII values shown as red data points; inner and outer lower fences indicated as black lines), whereas the ENCODE data displayed DNA methylation at this site for both cell lines (Fig 4.4 – MSRE_1 – red/orange methylated sites). Dynamic methylation was seen at the intragenic site that colocalised with H3K27ac and H3K4me3 enrichment (Fig 4.6B - MSRE_3: CT MspI shown in yellow, dCT HpaII values shown as green data points; inner and outer lower fences indicated as black lines) for A549 and Jurkat within the MSRE assay, whereas this is unmethylated for the Jurkat cell line and methylated for the A549 cell line within the ENCODE data (Fig 4.4 – MSRE_3 – green/blue for unmethylated sites and red/orange for methylated sites). Therefore, while our MSRE assay was consistent with an active, unmethylated enhancer at the *MuTE* site (Fig. 4.6, MSRE_5), there was variation between our results and those done in other laboratories, suggesting variation in cell lines and DNA methylation between laboratories.

| Table 4.8. DNA methylation statuses of Jurkat parental and clonal cell lines (P0, C11 and C4) as found by the MSRE assay for MSRE regions 2-5 and *TAL1* expression at passages 1 and 9.

	Region	2	3	4	5	*TAL1* Expression
Passage 1	P0	Unmethylated	Partial	Partial	Unmethylated	1.43
	C11	Unmethylated	Partial	Unmethylated	Unmethylated	0.75
	C4	Partial	High	Partial	Unmethylated	0.35
Passage 9	P0	Partial	High	Partial	Unmethylated	1.77
	C11	Partial	Partial	Partial	Unmethylated	0.58
	C4	Unmethylated	Partial	Partial	Unmethylated	0.40

[1] Methylation status was determined by boxplot outlier tests (Methods: 4.2.4) (Fig 4.6A).
[2] *TAL1* fold change relative to the Jurkat parental P0 at passage 1 as found in Chapter 2.

Overall, the results obtained from the MSRE assay for each cell line and passage, suggested an association with *TAL1* expression despite no statistical correlations (data not shown). MSRE_3 (co-localised with TSS peaks and H3K27ac and H3K4me3 enrichment) exhibited the most dynamic DNA methylation across the regions tested, specifically for the Jurkat parental at passage 1 and 9 (Fig 4.6 and Table 4.8 – P0 passage 1 and 9). It was seen that a state of partial methylation shifted to high methylation at this site and this coincided with increased *TAL1* expression (Table 4.8 – 1.43 to 1.77, relative fold expression), and was also seen for MSRE_2 (downstream of TSS Peaks and H3K27ac and H3K4me3 peak enrichment) shifting from an unmethylated state to a partially methylated state (Table 4.8 – region 2). In clonal cell line 11, within the downstream enriched DHS and H3K27ac intergenic site (MSRE_4), a shift from an unmethylated state to a partially methylated state was observed and this coincided with decreased *TAL1* expression (Table 4.8 – 0.75 to 0.58, relative fold expression). However, clonal cell line 4 did not exhibit a correlation between DNA methylation status and *TAL1* expression, with constant expression of *TAL1* between passages, despite changes in methylation status between passage 1 and 9 (Table 4.8 – region 2 and 3).

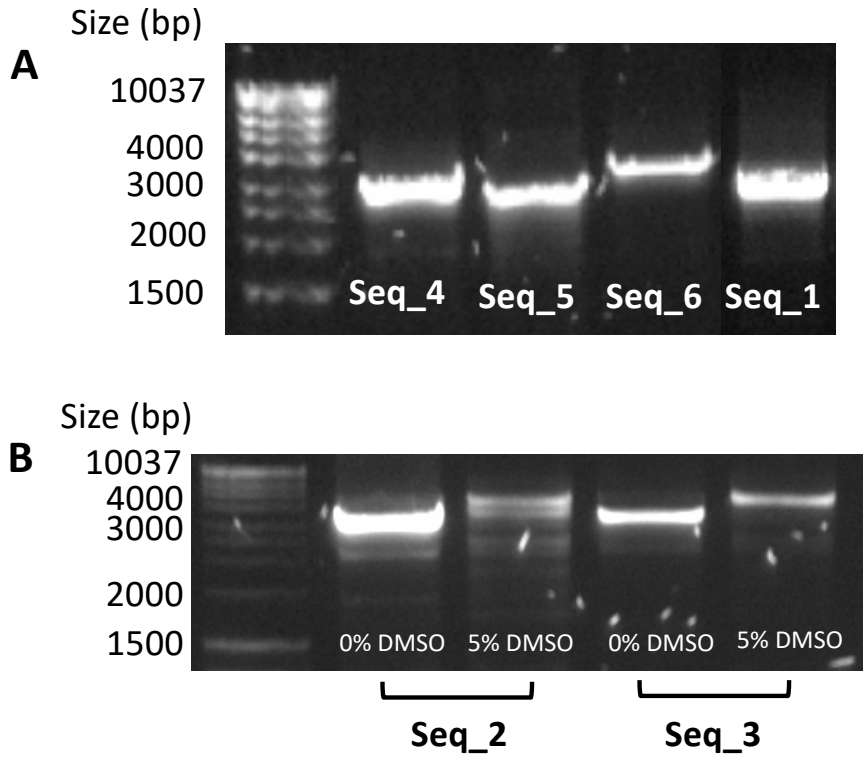
Using bioinformatics analysis, reduced HAIB representation bisulphite sequencing (RRBS), and HAIB Methylation450 array (Methyl450) data supplied from the ENCODE project, sites of DNA methylation within the *TAL1* locus were identified. Sites were identified that have the same or different patterns of DNA methylation for cell lines, such as A549 (lung epithelial cancer) and the Jurkat cell line. These sites have provided target regions for analysis of DNA methylation using the methylation-sensitive restriction endonuclease (MSRE) assay, as well as DNA sequencing using Nanopore sequencing of the Jurkat cell lines P0, C11 and C4 at passages 1 and 9.

4.3.4. Amplicon Generation Optimisation for Nanopore Sequencing

The initial cycling conditions suggested by New England BioLabs (Appendix – Supp. Table 7.28) were tested with all primers made for amplicon generation testing between different combinations of 5x Q5 High GC Enhancer at a 1x concentration, 50-200ng of gDNA, changes to the predicted annealing temperature (T_{AP}) (2°C + T_{AP}) and testing between 25µL and 50µL reactions. Each combination of the conditions tested provided no results when run on a 1.2% agarose gel stained with SYBR Safe Gel stain (data not shown). The GC content of the amplicons to be generated were then tested using the BiologicsCorp GC content calculator (https://www.biologicscorp.com/tools/GCContent) and provided a line graph to display the distribution of GC content across each amplicon (30bp window) (Appendix – Example shown in Supp. Fig 7.6).

After understanding the GC content of the amplicons to be generated, a new approach for the cycling conditions was taken place through the use of a touch-down PCR technique (Korbie and Mattick, 2008). A majority of primers approached GC content of 75-100% in areas of the amplicon to be generated (example – Supp. Fig 7.6). Therefore, as recommended by Korbie and Mattick (2008), 10% DMSO was used for amplicons with a GC-content > 60% in combination with the touch-down PCR technique with 75ng of gDNA with an annealing temperature of 10°C + T_{AP} and (10°C + T_{AP}) - 5°C (Initial annealing) and T_{AP} and T_{AP} - 5°C (second annealing temperature) with an extension time of 40 seconds/kb (Methods: 4.2.5). It was found that these conditions produced amplicons at the predicted size (2-4kb) for four out of the six primers for cycling conditions at 5°C below the T_{AP} (initial annealing temperature: (10°C + T_{AP}) - 5°C and second annealing temperature: T_{AP} - 5°C) (Fig 4.7A). The remaining two primers exhibited secondary PCR products at 10% DMSO, due to decreasing the melting point of the primers used (Lorenz, 2012). Therefore, conditions were tested for the remaining two primers at 0% and 5% DMSO, with secondary PCR

products and reduced amplification at 5% DMSO at the T_{AP} (Fig 4.7B – 5% DMSO). At 0% DMSO, increased single products at the predicted size were generated and therefore used for Jurkat cell line amplicon generation (Fig 4.7B - 0% DMSO).

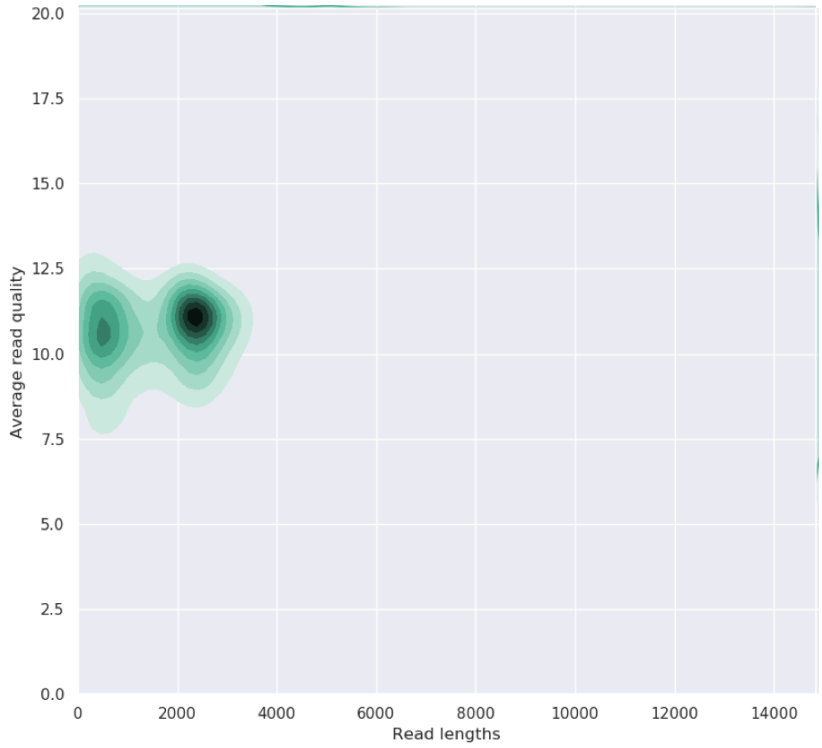


| Figure 4.7. (A-B). Optimisation of PCR amplicons (Seq_1 - Seq_6) with 75ng of gDNA using the touch-down PCR cycling conditions.

(Methods: 2.5.4). **A.** Amplicon primers Seq_1 and Seq_4 to Seq_6 under conditions of 10% DMSO with the T_{AP}-5°C touch-down PCR cycling conditions . **B.** Amplicon primers Seq_2 and Seq_3 optimised with 0% and 5% DMSO at the predicted annealing temperature (T_{AP}). At 5% DMSO, primers generated more secondary PCR products in comparison to 0% DMSO and therefore was used for Jurkat cell line amplicon generation.

After the sequencing run was completed, the MinION interface provided metrics on the quality, read lengths, cumulative output of the sequenced data and summary of channel pores used throughout the run of all the samples tested as a pool. This confirmed that the overall quality of the sequenced data had a q-score of 10 and approximately 100% of pores were used during the sequencing run (data not shown).

The quality of the data after processing and aligning (Methods: 4.2.7) (Aligned BAM files) was further validated per sample using Nanoplot within the Galaxy European database (https://usegalaxy.eu/). This provided similar metrics per sample and visualisation of the read quality displayed with a q-score > 7.5 (minimum 7 as stated by Theuns et al., (2018)) and a read length metric of 0-4kb (as expected for the amplicons generated) (Fig 4.8).



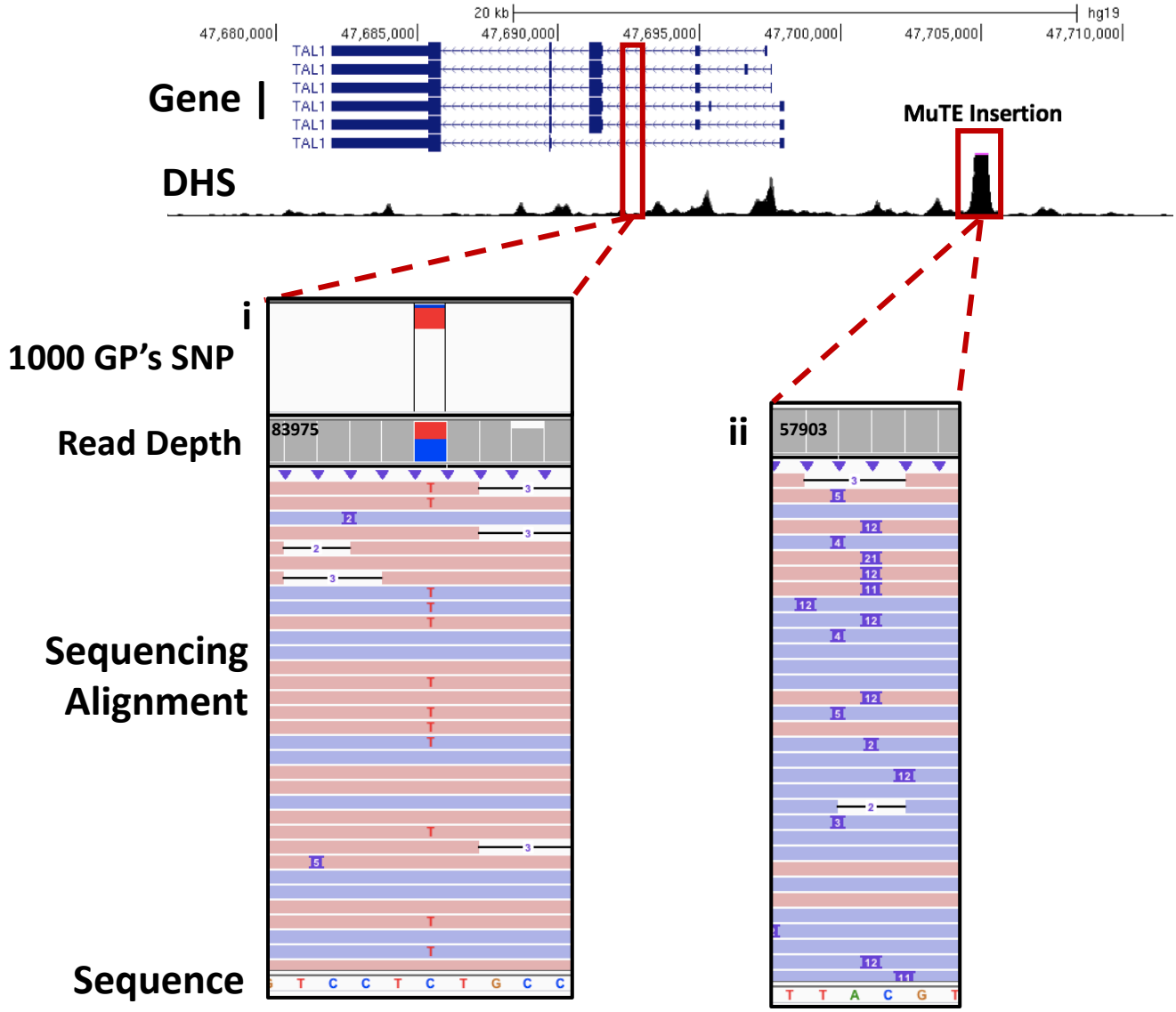
| Figure 4.8. Example of the metrics provided by European Galaxy database 'Nanoplot' for nanopore sequencing data.

The density plot displays the relationship between average read quality (y-axis) and the average read lengths of the sequences for sequencing of the Jurkat cell line amplicons, P0 at passage 1, across the *TAL1* locus. Density plot displays the correlation and concentration of values for average read quality across a q-score range of 7.5-12.5 and the average read length at two points around 0-1500bp and 2000-3500bp (range of amplicon sizes). Histogram plots along the right and top side of the plot display the scale of average read quality and read lengths individually to display the distribution of data (respectively).

The sequencing data was processed and assessed for genetic variants (Single nucleotide variants (SNVs)) using the 'Bcftools' mpileup file generated (Methods: 4.2.7). This identified genetic variants that differed from the reference genome (Human Genome Assembly GRCh37/hg19). Initially, between 20 to 27 genetic variants were called using 'Bcftools' per Jurkat cell line at each passage. Using the SP metric, genetic variants found per sample were narrowed down to 8-16 SNVs. A threshold of 60 was set (as stated by the Broad Institute) to remove a majority of false positive variants, however false positive variants can still be found within variants with a SP of < 60.

4.3.5. Nanopore sequencing of the Jurkat parental and clonal cell lines across *TAL1*

The Integrative Genomics Viewer (IGV) identifies variants using algorithms similar to the commonly used Genome Analysis Tooklit (GATK, Thorvaldsdóttir et al., 2013). Therefore, IGV was used to visually validate filtered SNVs identified by 'Bcftools' and provide allele frequencies of the called variants as well as highlight mismatched reads, highlight low quality reads and sort, group and colour alignments based on different parameters (read strand, base pair etc.). IGV also can map known GWAS and common SNPs from the 1000 Genomes Project, which in conjunction with the gnomAD database (Karczewski et al., 2019) was used to validate SNVs and other alleles in Jurkat cell line DNA across the *TAL1* locus (Example shown in Fig 4.9- i). Through using IGV, SNVs detected by 'Bcftools' were confirmed for 9 variants across all Jurkat cell lines, and one variant for clonal cell line 4 at passage 1 and 9 (Table 4.9). Five of the SNVs identified in Jurkat cell lines corresponded to known GWAS and common SNPs from the 1000 Genomes Project and the gnomAD database (Table 4.9, Fig 4.9 – i, see also Fig 4.10B). False-positive calling of insertions and deletions (indels) was seen using 'Bcftool' indel calling procedures, therefore IGV was used to manually find



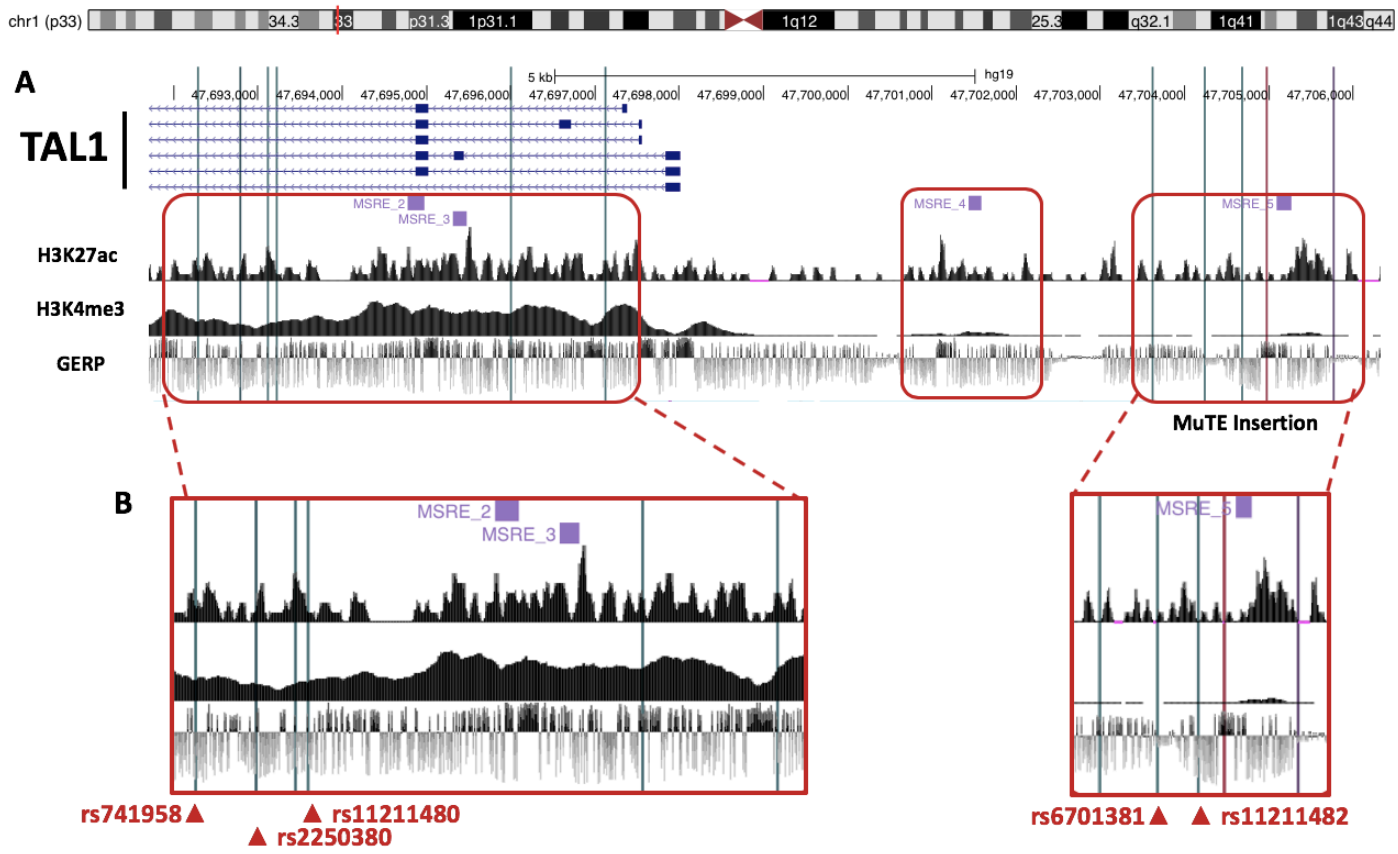
| Figure 4.9. IGV detection of a SNV and 12-bp insertion after Nanopore Sequencing of *TAL1*.

Top: The *TAL1* (Ref-seq) regulatory landscape using DNase1 Hypersensitivity data for Jurkat from the ENCODE project as displayed in the UCSC genome browser. **i).** Heterozygous 1000 Genomes Project SNP rs741958 at chr1: 47,692,281 in the Jurkat parental cell line (P0). Read-depth of sequencing is displayed (a total of 83975 reads that includes forward and reverse strands) with C: 56% and T: 43%. Alignment of nanopore sequencing is visualised in IGV with corresponding colours to read strands (red: forward, blue: reverse), with near-equal reads each strand for each allele **ii).** The 12-bp heterozygous *MuTE* insertion (chr1: 47,704,968), symbolised by purple marker with the insertion size displayed. Read depth of sequencing run is shown (57903) with alignment visualising of strands (red: forward, blue: reverse) and co-localising reference genome sequence relative to the insertion site. Approximately 50% of reads were assigned the insertion. Although most insertions were 12-bp in length, some sequencing-dependent artefacts of shorter or longer reads were also detected.

indels amongst Jurkat samples (Fig 4.9-ii). Indels were verified based on read depth and number of reads for each indel. Through manual screening of Jurkat samples using IGV, the 12-bp heterozygote insertion (predicted *MuTE* Insertion) was identified (Fig 4.9-ii purple bars, 50% of reads were assigned to the insertion) as stated in literature (Table 4.9 –Insertion) (Mansour et al., 2014).

Bioinformatic and MSRE analysis of the *TAL*1 locus in Jurkat T-cells (Chapters 3 and 4) identified three regulatory regions as defined by peaks of H3K27ac and/or H3K4me3 and sites of DNA methylation (Fig. 4.10A, red boxes). After confirmation of Jurkat cell line SNVs and the 12-bp insertion using IGV, each was mapped to the *TAL*1 locus: six variants mapped to introns within the *TAL*1 gene and flanked the MSRE sites MSRE_2 and MSRE_3 (Fig 4.10A, leftmost red box), while five genetic variants mapped to an intergenic region upstream of *TAL1* that also included the *MuTE* 12-bp insertion mutation and MSRE_5 (Fig. 4.10A, rightmost red box) No genetic variants mapped to the region that included MSRE_4 (Fig. 4.10A, middle red box).

After mapping the Jurkat SNVs within these regulatory regions, we next mapped co-localising genome-wide association studies (GWAS) and common SNPs from the 1000 Genomes Project and the gnomAD database (Karczewski et al., 2019; The 1000 Genomes Project Consortium, 2015; Zheng-Bradley and Flicek, 2017). It was found that 5 of the SNVs identified here localised with published SNPs (Fig 4.10B). Specifically, SNV1 and SNV2 mapped to GWAS SNPs rs741958 and rs2250380 that are associated with coronary artery calcified atherosclerotic plaques within African-Americans with type-2 diabetes (Divers et al., 2017); SNV4 mapped to rs11211480 that is associated with variation in human blood cell traits for complex disease (Astle et al., 2016); SNV8 and SNV9 mapped to published SNPs rs6701381 and rs11211482 that have not been reported in any GWAS study.



| Figure 4.10. Mapping of Jurkat cell line genetic variants within the *TAL1* locus identified using nanopore sequencing and co-localising published SNPS.

A. Location of Jurkat cell line genetic variants identified through 'Bcftools' and IGV as vertical lines across the *TAL1* landscape relative to regulatory element markers, H3K27ac and H3K4me3, MSRE primers and corresponding differential binding sites identified in Chapter 3 (red boxes). Chromosome ideogram of the *TAL1* locus and the Ref-Seq isoforms for *TAL1* are displayed. Vertical lines show the location of Jurkat cell line SNVs and Indels identified: dark green (1-9) – SNV found in all Jurkat cell lines tested, red (12) – Jurkat *MuTE* insertion identified in all Jurkat cell lines tested, dark purple (11) – SNV identified in C4 at passage 1 and 9 - numbers correspond to Table 4.9). GERP scoring across the *TAL1* locus displayed mammalian evolutionary conservation where positive scores (positive bar) display conserved bases and negative scores (negative bar) display evolutionary neutral bases. **B.** Zoomed-in view of the two regulatory regions that include Jurkat genetic variants identified in this study. *'rs'* numbers identify published SNPs from the 1000 Genomes Project and the GnomAD database and co-localise with SNVs 1, 2, 4, 8 and 9 (see also Table 4.9).

| Table 4.9. Called genotypes for finalised Jurkat cell line SNVs and Indels with a list of co-localising GWAS/common SNPs from 1000 Genomes Project and gnomAD database and their alternative allele and allele frequencies.

Jurkat Cell Line	SNV Coordinate (Chrom 1)	SNV Genotype	Co-localising GWAS[1]/common SNP[2]	SNP Alternative Allele – Allele[3] Frequencies
All cell lines	SNV1 - 47,692,281	Heterozygote	rs741958	T – 0.88
	SNV2 - 47,692,786		rs2250380 (also gnomAD identified SNP)	C – 0.88
	SNV3 - 47,693,116		-	-
	SNV4 - 47,693,220		rs11211480	G – 0.28
	SNV5 - 47,695,997	Homozygote (Alt)	-	-
	SNV6 - 47,697,125		-	-
	SNV7 - 47,703,613	Heterozygote	-	-
	SNV8 - 47,704,240		rs6701381	G – 0.34
	SNV9 - 47,704,674		rs11211482	T – 0.38
C4 Passage 1 and 9	SNV10 - 47,705,764	Heterozygote	-	-
All cell lines	Insertion – 47,704,968	12-bp heterozygote insertion (predicted MuTE insertion)	-	-

[1]GWAS-associated SNPs from the 1000 Genomes Project – rs2250380 also found within the gnomAD database

[2]Common SNPs from the 1000 Genomes Project

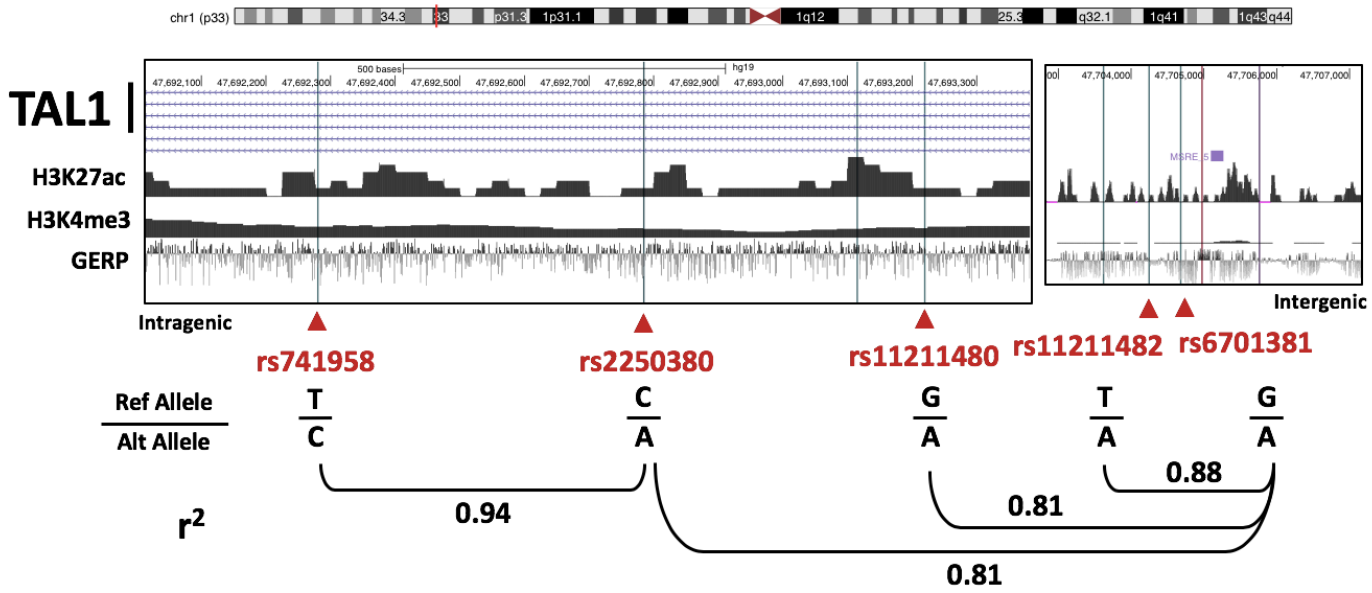
[3]Allele frequencies calculated from 1000 Genomes Project data

[4]All SNPs were identified as eQTLs from the GTEx database.

We next used the published SNPs that map to Jurkat SNVs to understand the genetic structure of the *TAL1* intragenic (SNV1, SNV2 and SNV4) and intergenic (SNV8 and 9) regions.

Linkage disequilibrium (LD) is the non-random association of alleles on the same chromosome (two or more loci) within a population (Slatkin, 2008). The human genome includes an average of 4 million SNPs and are identified through GWAS to map haplotype blocks across the genome (Corradin et al., 2014). A haplotype block is a genomic locus that is bound by common points of recombination and SNPs that fall within these blocks are seen to be in high LD (Wall and Pritchard, 2003). To identify haplotypes across the *TAL1* locus the NCBI LDLink database (Methods: 4.2.7) was used to find the 5 SNPs that co-localised with Jurkat SNVs 1, 2, 4, 8 & 9 (Fig 4.10B and Appendix – Supp. Table 7.29). It was found that LD extended across and between intragenic and intergenic regions, particularly seen with SNV9 (rs6701381) having high LD with SNV8/rs11211482 (r^2=0.88), SNV4/rs11211480 (r^2=0.81), and SNV2/rs2250380 (r^2=0.81) (Fig 4.11 – rs6701381). This analysis showed that all Jurkat SNVs 1, 2, 4, 8 & 9 are in high LD within a haplotype block. We also concluded that the remaining Jurkat SNVs that did not map to published SNPs, were novel mutations within this haplotype block.

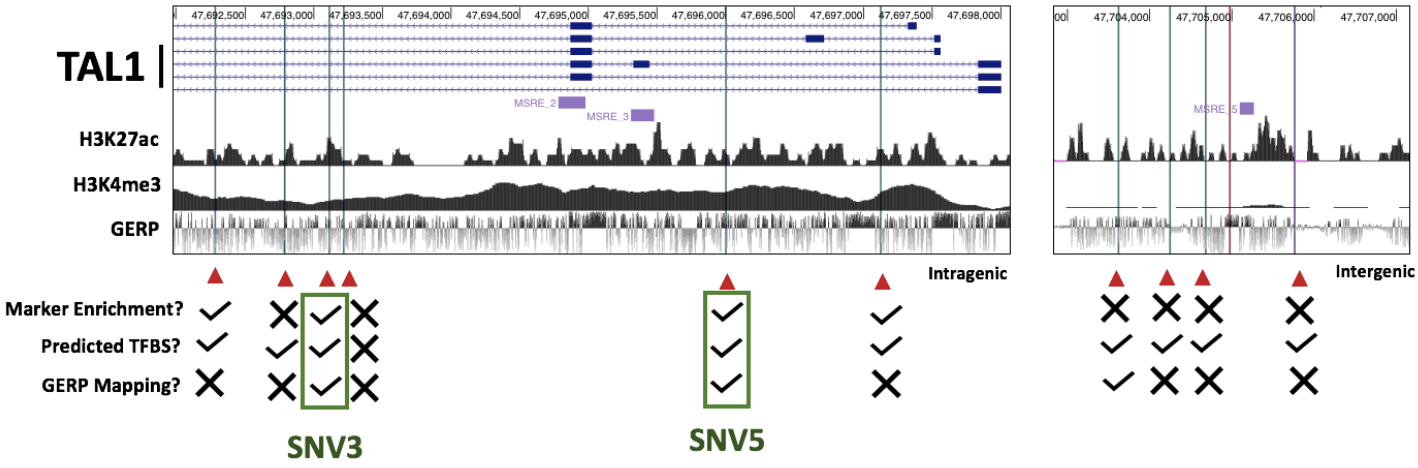
Having mapped the location of the Jurkat SNVs to two regulatory regions (Fig 4.10B) and identified patterns of LD across the *TAL1* locus (Fig 4.11), we next considered the possible functional relevance of the Jurkat SNVs in the context of regulatory elements such as promoters and enhancers. To do this, criteria that identify Jurkat SNVs predicted to have possible regulatory function were defined as: location within a regulatory element marker peak (H3K27ac and/or H3K4me3), predicted TF binding site(s) within 10bp upstream and downstream of the SNV, and Genomic Evolutionary Rate Profiling (GERP) localisation (Fig 4.12 and Appendix – Supp. Fig 7.7 to 7.16).



| Figure 4.11. Linkage disequilibrium patterns of co-localising published SNPs with Jurkat SNVs identified through Nanopore Sequencing within the *TAL1* locus.

Display of linkage disequilibrium (LD) patterns within intragenic and intergenic *TAL1* regions isolated from Fig 4.12 and zoomed in to display co-localising published SNPs with Jurkat SNVs. SNPs are shown to map to the associated variant and display the reference and alternative alleles for each SNV (all SNVs display heterozygous alternative alleles). LD correlation coefficient r^2, displays relationships of the co-localising SNPs within high LD (r^2>0.8). rs741958 has high LD with rs2250380 (r^2=0.94), rs2250380 also has high LD with rs6701381 (r^2=0.81), and rs6701381 has high LD with rs11211482 and rs11211480 (r^2=0.88 and 0.81, respectively).

After selection of SNVs based on these criteria, two SNVs (SNV3 and SNV5) within the intragenic region that do not map to published SNPs were identified. Specifically, SNV3 maps in the centre of a peak of H3K27ac and lower relative levels of H3K4me3, within a region of GERP identified high conservation with predicted binding of TFs, ZNF143, PKNOX2, TGIF1, TGFI2, NFE2 and NFE2 within 10bp (Fig 4.12 and Appendix – Supp. Fig 7.14). SNV5 also maps within a H3K27ac peak and H3K4me3,



| Figure 4.12. Prediction of Jurkat SNV functional relevance within regulatory elements.

The prediction analysis is based on the criteria of co-localising regulatory marker enrichment (H3K27ac and H3K4me3), predicted transcription factor binding from the JASPAR 2020 database within 10bp upstream and downstream of SNV and mapping of Genomic Evolutionary Rate Profiling (GERP) regions of conservation amongst mammalian species. (see also Appendix – Supp. Fig 7.7-7.16 for mapping of these criteria per SNV). JASPAR 2020 predicted sites display scoring of TF binging probability within a minimum p-value of $<10^{-4}$ (as stated by the JASPAR 2020 track in the UCSC genome browser). GERP scoring at each SNV, displays mammalian evolutionary conservation where positive scores (positive bar) display conserved bases and negative scores (negative bar) display evolutionary neutrally conserved bases.

with high predicted evolutionary conservation and predicted binding of TFs, EMOES and TBX21 (Fig 4.12 and Appendix – Supp Fig. 7.12). This suggested that the *TAL1* locus in the Jurkat cell lines included two rare variants/mutations that may play a role in regulatory element activity based on these criteria (Fig 4.12).

Having identified two possible regulatory SNVs, we next conducted a deeper predictive analysis of TF binding by using the JASPAR 2020 database (Khan et al., 2018b). Possible differences in TF binding to motifs that included alternate alleles of the two SNVs mutations were investigated by using 810 TF profiles for *Homo Sapiens* and within a range 12bp upstream and downstream of each variant allele, based on TF binding occurring across 6-12bp (Tuğrul et al., 2015). A comparison between TF

binding of the reference allele and the alternative allele for SNV3 and SNV5 was conducted. SNV3 displayed 27 predicted binding profiles for the reference allele, in comparison to 36 binding profiles with the alternative allele (G>A heterozygote) (Appendix – Supp. Table 7.30). In particular, predicted binding of TFs, MYB and NKX2-2 which are involved in T-ALL processes such as the TAL1 CRC and early cortical T-ALL (respectively)(Sanda et al., 2012; Van Vlierberghe and Ferrando, 2012) were identified within the reference G-allele analysis, but were absent for the alternate A-allele (Appendix – Supp. Table 7.30). Other TFs, such as SOX and HOXA proteins, were identified only for the alternate-A allele, and these TFs are associated with transcriptional regulation of embryonic and HSCs, respectively (Dou et al., 2016; Kamachi and Kondoh, 2013). For SNV5, predicted binding of the reference allele showed 30 TF binding profiles in comparison to 80 binding profiles for the homozygote alternative allele (G>A homozygote) (Appendix – Supp. Table 7.31). Of note, TFs GATA5, GATA2 and GATA3 were predicted to bind to a sequence that included the reference G-allele but were predicted to interact with increased affinity when tested with the alternate A-allele (e.g. GATA2: binding score of 4.45 to 5.10 and GATA3: 4.7 to 5.116) (Appendix – Supp. Table 7.31). The GATA3 TF is involved in TAL1 CRC function (Sanda et al., 2012), and similar to SNV3, the predicted binding of HOXA proteins was also seen with the alternate A-allele for SNV5.

4.4. Conclusion

A bioinformatics analysis using ENCODE supplied RRBS and Methyl450 data, mapped DNA methylation across the *TAL1* locus for the Jurkat cell line, as well as other leukaemia cell lines. This analysis showed that the *TAL1* locus was hypomethylated in Jurkat T-ALL cells and was consistent with published literature.

The MSRE assay was used to test the methylation landscape amongst Jurkat populations (P0, C11 and C4) at passages 1 and 9. Differential DNA methylation was seen between the different MSRE sites (exonic, two H3K27ac and H3K4me3 enriched sites, downstream intergenic H3K27ac and DHS site and the *MuTE* insertion enhancer site). Analysis of the intragenic MSRE sites that co-localised with TSS peaks, H3K27ac and H3K4me3 showed some Jurkat populations displaying high methylation dependent on passage (C4 and P0), while the intergenic Jurkat (*MuTE*) enhancer showed all Jurkat populations were unmethylated at this site for all passages, as expected for tumorigenic *TAL1* overexpression. Trends were seen between DNA methylation status within the parental cell line, specifically, intragenic methylation increased as *TAL1* expression increased, whereas clonal cell line 11 showed that DNA methylation increased at the downstream intergenic H3K27ac and DHS peak with a decrease of *TAL1* expression at passage 1 and 9 (Table 4.8, MSRE_2). This identified two regions of potential dynamic DNA methylation that were associated with changes in gene expression patterns between Jurkat cell lines and passages, and which might be consistent with the presence of an intragenic enhancer (King et al., 2016; Mansour et al., 2014), particularly for MSRE_3. Therefore, an investigation into possible genetic differences across the *TAL1* locus between these Jurkat cell lines could determine why differences in the phenotype, transcription of *TAL1* and the DNA methylation profile within the *TAL1* locus are seen. To do this, a Nanopore sequencing experiment was conducted for parental and clonal populations P0, C11 and C4 at passages 1 and 9.

The sequencing of the *TAL1* locus identified genetic variants within Jurkat parental and clonal cell lines, P0, C11 and C4, at sites of dynamic DNA methylation (MSRE_2 and 3) and intragenic sites enriched for H3K27ac and H3K4me3 as well as an intergenic site enriched for H3K27ac. Five of these variants identified also mapped with publicly available GWAS and common SNP data from the 1000 Genomes Project

and the gnomAD database, allowing us to conduct a predictive analysis of recombination patterns between these SNPs. These SNPS were found to be in high LD with each other, suggesting they all fall within one haplotype block that extended between intragenic and intergenic regions. The genetic variants were further analysed for their specific mapping to regulatory element markers, highly conserved regions and potential TF binding. This identified two SNVs that met these criteria, SNV3 and SNV5, which were further analysed using the JASPAR database for predicted binding of TFs when comparing the reference and alternative allele within the sequence. For both SNVs, the presence of the alternative allele is correlated with a difference in the predicted number, identity and affinity of TFs binding to motifs that include these alleles. However, SNV3 was seen to no longer contain MYB and NKX2-2 binding motifs after alteration of the allele but introduced SOX and HOXA protein binding. The alternative allele for SNV5 was associated with increased predicted TF binding such as for the key TAL1 CRC protein GATA3 and other HOXA proteins as well. Therefore, these two SNVs present in all cell lines may be novel variants that are predicted to be associated with regulatory elements within the *TAL1* locus but are not Jurkat cell-line specific.

Chapter 5 - Discussion and Future Directions

5.1. Hypothesis

The hypothesis of this thesis proposed that the expression of the gene encoding the transcription factor TAL1 was associated with the proliferation of the Jurkat late cortical T-ALL cell line. Specifically, we tested whether differences in the proliferation of Jurkat T-ALL clonal cell lines reflected differences in *TAL1* transcription, and whether such differences were accompanied by epigenetic and/or genetic differences between the different cell lines and at varying passage numbers of culture.

5.2. Summary of Results

5.2.1. Phenotypic and Transcriptional Profiles of Jurkat parental and clonal cell lines

In this thesis we have shown that Jurkat clonal populations have differential proliferative abilities that are typically higher than the original parental population (Chapter 2 – Fig 2.7). Differential proliferation was statistically significant between clonal and parental populations, as clonal cell lines tended to show higher rates of proliferation with increasing time in continuous culture (Chapter 2 – Fig 2.7 and Table 2.3). This was expected as Ben-David et al. (2018) and Zhao et al. (2012) reported continual passaging altered the proliferative qualities of cancer cells and stem cells respectively. We also tested the expression of *TAL1*, *GATA3*, *RUNX1* and *MYB* (all the members of the TAL1 CRC), which showed a general trend of decreasing expression of *TAL1* in the clonal cell lines relative to the parental cell lines, with the parental cell lines falling within the *TAL1*$^{+}$ expression group (Chapter 2 – Table 2.5). This was not expected as both parental cell lines displayed a lower median proliferative value

relative to the isolated clonal populations, with the majority of clonal cell lines falling within the extreme proliferation group (Chapter 2 – Table 2.5). Contrary to our proposed hypothesis, these results showed an inverse relationship between *TAL1* expression and proliferation (Chapter 2 – Fig 2.7 and Table 2.5) (Litzow and Ferrando, 2015). Ben-David et al. (2018) also demonstrated difficulty in establishing a clear link between differential drug responses of cancer cell lines tested and proliferation rate. This suggests that proliferation of Jurkat T-ALL cells may not be as tightly controlled by the TAL1 CRC as expected. Overall, in agreement with earlier studies by Ben-David et al. (2018) and Martin-Pardillos et al. (2019), we have demonstrated phenotypic and transcriptional heterogeneity between Jurkat clonal populations and the original parental populations they were derived from, showing decreased *TAL1* expression associated with hyper-proliferative clonal populations.

5.2.2. The understanding of the intergenic *MuTE* insertion enhancer within Jurkat cell lines

Using publicly available data from the ENCODE consortium (Chapter 3 – Fig 3.4), we next confirmed the location of the powerful *TAL1* Jurkat super-enhancer that maps to a predicted *MuTE* insertion (Mansour et al., 2014; Tan et al., 2019), with H3K27ac enrichment symbolising this super-enhancer and spreading into the downstream *TAL1* gene. This *in silico* analysis of an active super-enhancer agreed with our *in vitro* analysis of this region which revealed stable DNA hypomethylation for all Jurkat cell lines, and supported the idea that the *MuTE* intergenic enhancer increases *TAL1* expression when hypomethylated (Chapter 4 – Table 4.8) (Ehrlich et al., 2016; Rauscher et al., 2015). The Mansour et al. (2014) study also shows that the *MuTE* super-enhancer is essential for Jurkat T-ALL cell viability as deletion of this region using CRISPR/Cas9 gene editing failed to produce any viable cells. Our identification of the 12-bp heterozygous insertion at the *MuTE* site through nanopore sequencing in all cell lines tested would support the idea that this insertion is important for Jurkat T-ALL

cell viability. Overall, confirming the location, activity and importance of the *MuTE* insertion amongst Jurkat cell lines.

5.2.3. The intragenic landscape of *TAL1* within Jurkat parental and clonal cell lines

Through bioinformatic analysis, two intragenic sites were identified within the first two introns of *TAL1* that map to statistically significant H3K27ac enrichment between the Jurkat cell line relative to other non-immune cell lines and primary T-cells but not to the DND41 cell line (non-TAL1 T-ALL cell line) (Chapter 3 – Fig 3.4). This predictive analysis agrees with the findings by Mansour et al. (2014), who show that the heterozygous somatic mutation within the non-coding intergenic region of *TAL1*, creates a super-enhancer with H3K27ac enrichment breadth extending into the first introns of the *TAL1* gene within Jurkat cell lines. In addition, active promoter marker H3K4me3 enrichment also mapped to this intragenic region and was statistically significant for the Jurkat cell line but was not assessed within the Mansour et al. (2014) study. This suggested putative promoter activity (Calo and Wysocka, 2013) (Chapter 3 – Fig 3.4) however, mapped RNA-seq transcripts from the GTEx database for thyroid and whole-blood cell types suggested that the main promoter activity was at the 5' end of the *TAL1* gene as stated by Patel et al. (2014) (Chapter 3 – Fig 3.7). Therefore, we conclude the existence of an intragenic regulatory element at this site within the Jurkat cell line.

We have argued that the presence of H3K27ac and H3K4me3 enrichment and the evidence of weak transcriptional activity at this intragenic site has suggested the presence of an intragenic regulatory DNA element. We analysed the methylation status of this site using the MSRE assay (MSRE_2 and MSRE_3). DNA methylation at MSRE_3 which maps to the intragenic H3K27ac and H3K4me3 enriched peak displayed high DNA methylation within the P0 cell line despite DNA methylation at

CpG sites elsewhere within *TAL1* displaying DNA hypomethylation (Chapter 4 – Fig 4.7 and Table 4.6) (Haider et al., 2018; Vandiver et al., 2015). The high DNA methylation of this site in Jurkat parental populations contrasts with reduced methylation in clonal cell lines 11 and 4, and suggested this region showed dynamic DNA methylation (Chapter 4 – Table 4.8). The higher levels of *TAL1* gene expression in the parental cell line would be consistent with high intragenic methylation being associated with high levels of gene transcription (Neri et al., 2017; Rauscher et al., 2015). We would argue that this intragenic region is an intragenic enhancer, with high methylation inhibiting intragenic enhancer-associated transcription that would otherwise inhibit *TAL1* expression (Cinghu et al., 2017; Varley et al., 2013) (Chapter 4 – Table 4.8). Dynamically methylated regions that map to regulatory elements have been found to be at tissue-specific regions associated with key cell identity genes that require dynamic switching that results in cell-to-cell heterogeneity (Song et al., 2019). TAL1 is known to be involved in HSC tissue-specificity, therefore aligns with why DNA methylation would be dynamic at this region and lead to heterogeneity within T-ALL as seen between the Jurkat cell lines (Sanda et al., 2012; Song et al., 2019).

Finally, MinION nanopore sequencing identified nine SNVs that mapped within the *TAL1* locus at DBA identified intragenic and intergenic sites common amongst all Jurkat cell lines (P0, C11 and C4 at passage 1 and 9), except for an additional heterozygote variant within the clonal cell line 4 at passage 1 and 9 (Chapter 4 – Table 4.9 - SNV10). Though Jurkat cell line mutations were expected due to the known instability of the cell line as stated by Gioia et al. (2018), our hypothesis also proposed cell-line specific SNVs, which was only seen for clonal cell line 4 upstream of the intergenic *MuTE* insertion. Five out of the 10 SNVs identified co-localised with published SNPs from the 1000 Genomes Project in which the five SNPs were in high LD with one another, suggesting a predicted haplotype block that all Jurkat SNVs mapped within for intragenic and intergenic regions of *TAL1*. Specifically, the

predicted *TAL1* intragenic enhancer identified by DBA and dynamic DNA methylation mapped two novel SNVs (SNV3 and SNV5) flanking MSRE_3 which were predicted for association with regulatory elements through predicted TF binding and high evolutionary conservation at these sites (Chapter 4 – Fig 4.12). TF binding motif analysis of SNV3 and SNV5 (12-bp upstream and downstream) identified differences in the predicted TF binding profiles when comparing the reference allele and the alternative alleles for each SNV within a 24-bp sequence. This agrees with our hypothesis that genetic variants in cancer cells map to regulatory DNA elements of *TAL1*, consistent with increasing evidence identifying functional SNVs in the non-coding segments of the genome (Scacheri and Scacheri, 2015). SNV3 and SNV5 investigations into the predicted binding of TFs at the reference and alternative allele displayed increased differences in the number and affinity of TF binding for the alternative allele, such as HOXA and SOX TFs (Appendix – Table 7.30 and 7.31). HOXA and SOX proteins are known for their association to embryogenic and HSC transcriptional activity (Dou et al., 2016; Kamachi and Kondoh, 2013), in which HOXA TFs are also known as master TFs within T-ALL, and are associated with HSC differentiation inhibition when overexpressed, similar to aberrant TAL1-mediated T-ALL programs (Bond et al., 2016; Tan et al., 2019). Overall, this has displayed the possible introduction and increased affinity of TFs that aberrantly bind at the predicted intragenic enhancer of *TAL1* with possible relevance to T-ALL.

Thus, we have demonstrated evidence that the intragenic region of the *TAL1* locus is a modulated intragenic enhancer when considering predicted regulatory activity within this site using DBA, through the dynamic DNA methylation profile identified by the MSRE assay, the correlation of increased DNA methylation at this intragenic site and increased *TAL1* expression and the mapping of potential regulatory-associated SNVs for all the Jurkat cell lines at this site.

5.3. Future Directions

In Chapter 2, it was difficult to see the distinct differences between proliferation of each individual clonal population and a clear correlation with gene expression of TAL1 CRC genes. Consequently, relationships between proliferation and gene expression of late cortical TAL1 CRC genes may not be apparent due to only focusing on a small subset of possible genes (TAL1 CRC) that may be partly involved or not involved in TAL1-mediated T-ALL (Girardi et al., 2017; Van Vlierberghe and Ferrando, 2012). In addition, conducting serum starvation and serum re-introduction as a treatment to initiate proliferation may not have drawn out the differences between the clonal populations to a point of statistical significance and was not reflected within the TAL1 CRC gene expression assay statistically as well. Therefore as a future direction, other treatments that can draw out more significant differences between clonal cell lines may be conducted such as T-ALL-specific microenvironment stimuli such as Interleukin 7 (IL7) and Sonic hedgehog (SHH) signalling is known to induce survival and proliferation within T-ALL cell types *in vivo* through NOTCH1 pathways (a gene involved in all T-ALL subtypes) and cortico-medullary subtypes such as TAL1-mediated T-ALL, respectively (Passaro et al., 2016). This may draw out statistical differences between Jurkat clonal cell lines that is more related to *in vivo* processes within the microenvironment. The analysis into proliferation can also be tested for relative gene expression with a larger set of genes involved in late cortical T-ALL proliferation and activity such as late-cortical T-ALL specific genes *LMO1*, *LMO2*, *TAL2* and other T-ALL general genes such as *NOTCH1*, *CDKN2A* and *CDKN1B* (Chapter 1 – Table 1.1), which are involved in cell cycle regulatory functions for a more specific gene approach (Girardi et al., 2017; Van Vlierberghe and Ferrando, 2012). However, genome-wide RNA-seq would be the benchmark technique to map transcripts that show a significant difference in expression within hyperproliferative

clonal cell lines and be correlated to proliferative phenotypes for a more large-scale robust analysis. This has been a common technique utilised for identifying proliferation-informative cancer types through screening thousands of genes to develop a proliferation-specific gene signature (Ramaker et al., 2017; Waldman et al., 2013).

The analysis conducted in Chapter 3 confirmed the location and specificity of the *MuTE* insertion super-enhancer within the Jurkat cell line, however it is not known whether the other enhancer elements found within the analysis are a part or work in conjunction with the enhancer found at this site, such as the intergenic H3K27ac peak downstream of the *MuTE* insertion enhancer (Chapter 3 - Fig 3.4 - intergenic downstream enhancer peak). These elements identified across the *TAL1* locus can then be further stratified by utilising the H3K4me2 marker to define "hyperacetylated chromatin domains" (HCDs) in conjunction with the active regulatory element marker, H3K27ac. These regions distinctly confirm and stratify super-enhancers that align more closely with cell type-specific activity and coordination of enhancer elements as a cohesive unit, as opposed to other super-enhancers that don't map H3K4me2 and may not function differently from typical enhancers (Fox et al., 2019). Furthermore, due to this bioinformatic analysis being predictive and utilising data from other studies, a display of regulatory element markers within our Jurkat clonal populations was not assessed and therefore may not truly reflect the actual functioning of regulatory elements within *TAL1* and the relevant isoforms within these clonal populations. Using ChIP-seq analysis of H3K27ac and H3K4me3 can be utilised within a differential binding analysis to statistically assess the activity of regulatory elements across the *TAL1* locus between clonal populations and the parental cell line (Brown, 2011; Steinhauser et al., 2016). Whilst utilising the RNA-seq technique can aid in identifying where isoform transcripts originate from within the *TAL1* locus to ensure the proper localisation of the TSS for the Jurkat cell lines used

within the project (Sarantopoulou et al., 2019; Zhang et al., 2017a). This will confirm if whether the intragenic region identified is a true promoter for *TAL1* within the Jurkat cell line (Chapter 3 – Fig 3.7).

In Chapter 4, the MSRE assay provided a focus of broad patterns of dynamic DNA methylation within *TAL1*, it has provided a cost-effective model to be used to assess other genes involved in TAL1-mediated pathogenesis such as the TAL1 CRC and other late cortical T-ALL related genes such as *LMO1*, *LMO2*, *NOTCH1* and *TAL2* (Melnikov et al., 2005; Van Vlierberghe and Ferrando, 2012). Due to the results of this analysis displaying dynamic and stable DNA methylation patterns within just *TAL1* between Jurkat cell lines, similar patterns specific to T-ALL are likely to be throughout entire Jurkat genomes. The MSRE assay allows for predictive testing before conducting quantitative costly-assays for DNA methylation such as RRBS, which would be the next step for assessing whole-genome DNA methylation patterns between Jurkat clonal cell lines tested in the context of late cortical T-ALL heterogeneity (Meissner et al., 2005). Other possible future directions would investigate whether the predicted intragenic enhancer site is a part of the *MuTE* insertion super-enhancer and test the relationship of super-enhancer activity and DNA methylation as conducted by Heyn et al. (2016). This may illuminate if super-enhancer activity at the *MuTE* insertion site may be conducted and modulated through individual enhancer elements or if the multiple subunits work as a cohesive unit, a phenomenon which is still unknown (Jia et al., 2019).

Finally, future directions using techniques such as ChIP-seq for H3K27ac and H3K4me3 enrichment and DNase-seq for DHS, can be used to assess the individual activity patterns of the regulatory elements identified between the different Jurkat clonal populations between passages 1 and 9, particularly the variability within C11 and the low proliferative and TAL1 expression within C4 (Chapter 2 – Table 2.3 and

Table 2.5). These techniques could also be used to assess TF binding of TAL1 CRC TFs as well such as TAL1, GATA3, RUNX1 and MYB, specifically for these sites and assess the functionality of the Jurkat SNVs directly across *TAL1* for TF binding affinity. Long-read nanopore sequencing can also be utilised to confirm the haplotype block which was predicted using published SNPs amongst late cortical T-ALL Jurkat cell lines, as it provides better distinction of haplotypes and the allele-specific expression that may co-localise with the regulatory elements identified in Chapter 3 (Ebler et al., 2019; Mantere et al., 2019). This analysis has also provided a workflow of testing gene-specific genetic variants between clonal populations that can be applied to many other TAL1-T-ALL related genes such as the TAL1 CRC. This may be in particular interest for C4 that falls within the *GATA3*$^{+}$ gene group for assessing if *GATA3* epigenetic variability and genetic variants that map *GATA3* are the reason for the relatively low proliferative abilities of the clonal cell line 4 relative to the other clonal cell lines between passages (Chapter 2 – Table 2.3 and Table 2.5). The heterogeneity presented between Jurkat clonal cell lines can also be assessed between single cells within each population to confirm whether cell populations within each clonal cell line harbour or contain the same transcriptional, epigenetic and genetic basis as this study was based on the average trends seen between whole populations. Techniques such as single-cell RNA sequencing or DNA sequencing can be used to assess these differences between single cells within each clonal cell line for whole genome profiles (Goldman et al., 2019; Singh et al., 2019; Sun et al., 2019). This technique has the ability to correlate and differentiate sequences within a given cell across proteomic, epigenomic, transcriptomic and genetic levels, adding depth into the analysis of intra-tumoural heterogeneity amongst Jurkat clonal cell lines (Goldman et al., 2019; Sun et al., 2019)

5.5. Conclusion

In conclusion, this thesis has provided a cost-effective model to generate a resource of clonal cell lines and provides evidence of potential mechanisms of heterogeneity displayed amongst the cell lines, specifically in the context of the *TAL1* locus. These cell lines can be exploited to further assess heterogeneity within populations in a cost-effective way and do functional testing to test other cancer hallmark phenotypes, such as the differentiation block within T-ALL, as well as test a variety of other genes that elaborate on the TAL1 CRC and its role in TAL1-mediated T-ALL. We have also established a rationale for testing DNA methylation across different T-ALL related loci that map within intragenic and intergenic enhancers, for a deeper understanding of the malleability of cancer genomes and its role in heterogeneity. This thesis has also provided evidence of novel Jurkat specific SNVs within the clonal populations that can be further tested for their relevance in the functionality of the regulatory elements, epigenetic sites and TF binding in late cortical T-ALL. Not only have we shown evidence for *TAL1* being a gene target that displays clonal cell line heterogeneity through identifying dynamic DNA methylation patterns at these sites, but this gene-specific approach can also be applied to the rest of the genome to understand the basis of late cortical T-ALL and other cancers.

Chapter 6 - References

Abraham, R.T., and Weiss, A. (2004). Jurkat T cells and development of the T-cell receptor signalling paradigm. Nat. Rev. Immunol. *4*, 301–308.

Akoglu, H. (2018). User's guide to correlation coefficients. Turk. J. Emerg. Med. *18*, 91–93.

Almamun, M., Kholod, O., Stuckel, A.J., Levinson, B.T., Johnson, N.T., Arthur, G.L., Davis, J.W., and Taylor, K.H. (2017). Inferring a role for methylation of intergenic DNA in the regulation of genes aberrantly expressed in precursor B-cell acute lymphoblastic leukemia. Leuk. Lymphoma *58*, 2156–2164.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415-1429.e19.

Azarsiz, E., Karaca, N., Ergun, B., Durmuscan, M., Kutukculer, N., and Aksu, G. (2018). In vitro T lymphocyte proliferation by carboxyfluorescein diacetate succinimidyl ester method is helpful in diagnosing and managing primary immunodeficiencies. J. Clin. Lab. Anal. *32*, e22216.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput. Biol. *9*.

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C.A., Dempster, J., Lyons, N.J., Burns, R., et al. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. Nature *560*, 325–330.

Benveniste, D., Sonntag, H.-J., Sanguinetti, G., and Sproul, D. (2014). Transcription factor binding predicts histone modifications in human cell lines. Proc. Natl. Acad. Sci. *111*, 13367–13372.

Bird, A., Tate, P., Nan, X., Campoy, J., Meehan, R., Cross, S., Tweedie, S., Charlton, J., and Macleod, D. (1995). Studies of DNA methylation in animals. J. Cell Sci. Suppl. *19*, 37–39.

Blinka, S., Reimer, M.H., Pulakanti, K., Pinello, L., Yuan, G.-C., and Rao, S. (2017). Identification of Transcribed Enhancers by Genome-wide Chromatin Immunoprecipitation Sequencing. Methods Mol. Biol. Clifton NJ *1468*, 91–109.

Bocharov, G., Luzyanina, T., Cupovic, J., and Ludewig, B. (2013). Asymmetry of Cell Division in CFSE-Based Lymphocyte Proliferation Analysis. Front. Immunol. *4*.

Bond, J., Marchand, T., Touzart, A., Cieslak, A., Trinquand, A., Sutton, L., Radford-Weiss, I., Lhermitte, L., Spicuglia, S., Dombret, H., et al. (2016). An early thymic precursor phenotype predicts outcome exclusively in HOXA-overexpressing adult T-cell acute lymphoblastic leukemia: a Group for Research in Adult Acute Lymphoblastic Leukemia study. Haematologica *101*, 732–740.

Bowden, R., Davies, R.W., Heger, A., Pagnamenta, A.T., de Cesare, M., Oikkonen, L.E., Parkes, D., Freeman, C., Dhalla, F., Patel, S.Y., et al. (2019). Sequencing of human genomes with nanopore technology. Nat. Commun. *10*.

Brown, S.R. (2011). DiffBind: differential binding analysis of ChIP-seq peak data.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. *55*, 611–622.

Bůžková, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene—gene and gene—environment interactions. Ann. Hum. Genet. *75*, 36–45.

Calo, E., and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? Mol. Cell *49*, 825–837.

Can, T. (2014). Introduction to Bioinformatics. In MiRNomics: MicroRNA Biology and Computational Analysis, M. Yousef, and J. Allmer, eds. (Totowa, NJ: Humana Press), pp. 51–71.

Carelli, F.N., Liechti, A., Halbert, J., Warnefors, M., and Kaessmann, H. (2018). Repurposing of promoters and enhancers during mammalian evolution. Nat. Commun. *9*.

Carlotti, E., Pettenella, F., Amaru, R., Slater, S., Lister, T.A., Barbui, T., Basso, G., Cazzaniga, G., Rambaldi, A., and Biondi, A. (2002). Molecular characterization of a new recombination of the SIL/TAL-1 locus in a child with T-cell acute lymphoblastic leukaemia. Br. J. Haematol. *118*, 1011–1018.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. *74*, 106–120.

Chang, B.H.W., and Tian, W. (2016). GSA-Lightning: ultra-fast permutation-based gene set analysis. Bioinformatics *32*, 3029–3031.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

Chiaretti, S., Gianfelici, V., O'Brien, S.M., and Mullighan, C.G. (2016). Advances in the Genetics and Therapy of Acute Lymphoblastic Leukemia. 9.

Cinghu, S., Yang, P., Kosak, J.P., Conway, A.E., Kumar, D., Oldfield, A.J., Adelman, K., and Jothi, R. (2017). Intragenic enhancers attenuate host gene expression. Mol. Cell *68*, 104-117.e6.

Clough, E., and Barrett, T. (2016). The Gene Expression Omnibus database. Methods Mol. Biol. Clifton NJ *1418*, 93–110.

Coccaro, N., Anelli, L., Zagaria, A., Specchia, G., and Albano, F. (2019). Next-Generation Sequencing in Acute Lymphoblastic Leukemia. Int. J. Mol. Sci. *20*.

Consortium, T.E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal·lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. *24*, 1–13.

Curtis, A.E., Smith, T.A., Ziganshin, B.A., and Elefteriades, J.A. (2016). The Mystery of the Z-Score. AORTA J. *4*, 124–130.

Danecek, P., and McCarthy, S.A. (2017). BCFtools/csq: haplotype-aware variant consequences. Bioinformatics *33*, 2037–2039.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., and Fuks, F. (2014). A comprehensive overview of Infinium HumanMethylation450 data processing. Brief. Bioinform. *15*, 929–941.

Deshpande, Reed, Sullivan, Kerkhof, Beigel, and Wade (2019). Offline Next Generation Metagenomics Sequence Analysis Using MinION Detection Software (MINDS). Genes *10*, 578.

Divers, J., Palmer, N.D., Langefeld, C.D., Brown, W.M., Lu, L., Hicks, P.J., Smith, S.C., Xu, J., Terry, J.G., Register, T.C., et al. (2017). Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes. BMC Genet. *18*, 105.

Dou, D.R., Calvanese, V., Sierra, M.I., Nguyen, A.T., Minasian, A., Saarikoski, P., Sasidharan, R., Ramirez, C.M., Zack, J.A., Crooks, G.M., et al. (2016). Medial HOXA genes demarcate haematopoietic stem cell fate during human development. Nat. Cell Biol. *18*, 595–606.

Ebler, J., Haukness, M., Pesout, T., Marschall, T., and Paten, B. (2019). Haplotype-aware diplotyping from noisy long reads. Genome Biol. *20*, 116.

Ehrlich, K.C., Paterson, H.L., Lacey, M., and Ehrlich, M. (2016). DNA Hypomethylation in Intragenic and Intergenic Enhancer Chromatin of Muscle-Specific Genes Usually Correlates with their Expression. Yale J. Biol. Med. *89*, 441–455.

Eisinga, R., Heskes, T., Pelzer, B., and Te Grotenhuis, M. (2017). Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers. BMC Bioinformatics *18*, 68.

ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046.

Fernández-Ramos, A.A., Marchetti-Laurent, C., Poindessous, V., Antonio, S., Petitgas, C., Ceballos-Picot, I., Laurent-Puig, P., Bortoli, S., Loriot, M.-A., and Pallet, N. (2017). A comprehensive characterization of the impact of mycophenolic acid on the metabolism of Jurkat T cells. Sci. Rep. *7*, 1–11.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database J. Biol. Databases Curation *2017*.

Fouad, Y.A., and Aanei, C. (2017). Revisiting the hallmarks of cancer. Am. J. Cancer Res. *7*, 1016–1036.

Fox, S., Lillis, J.A., Davidson, C., Getman, M., Kingsley, P.D., and Bulger, M. (2019). Hyperacetylated Chromatin Domains Mark Cell Type-Specific Genes and Suggest Distinct Modes of Enhancer Function. BioRxiv 666784.

Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503–508.

Gioia, L., Siddique, A., Head, S.R., Salomon, D.R., and Su, A.I. (2018). A genome-wide survey of mutations in the Jurkat cell line. BMC Genomics *19*, 334.

Girardi, T., Vicente, C., Cools, J., and De Keersmaecker, K. (2017). The genetics and molecular biology of T-ALL. Blood *129*, 1113–1123.

Glazko, G., and Mushegian, A. (2010). Measuring gene expression divergence: the distance to keep. Biol. Direct *5*, 51.

Goldberg, J.M., Silverman, L.B., Levy, D.E., Dalton, V.K., Gelber, R.D., Lehmann, L., Cohen, H.J., Sallan, S.E., and Asselin, B.L. (2003). Childhood T-Cell Acute Lymphoblastic Leukemia: The Dana-Farber Cancer Institute Acute Lymphoblastic Leukemia Consortium Experience. J. Clin. Oncol. *21*, 3616–3622.

Goldman, S.L., MacKay, M., Afshinnekoo, E., Melnick, A.M., Wu, S., and Mason, C.E. (2019). The Impact of Heterogeneity on Single-Cell Sequencing. Front. Genet. *10*.

Gutschner, T., and Diederichs, S. (2012). The hallmarks of cancer. RNA Biol. *9*, 703–719.

Haider, Z., Larsson, P., Landfors, M., Köhn, L., Schmiegelow, K., Flaegstad, T., Kanerva, J., Heyman, M., Hultdin, M., and Degerman, S. (2019). An integrated transcriptome analysis in T-cell acute lymphoblastic leukemia links DNA methylation subgroups to dysregulated *TAL1* and ANTP homeobox gene expression. Cancer Med. *8*, 311–324.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hashimoto, K., Kokubun, S., Itoi, E., and Roach, H.I. (2007). Improved Quantification of DNA Methylation Using Methylation-Sensitive Restriction Enzymes and Real-Time PCR. Epigenetics *2*, 86–91.

Heyn, H., Vidal, E., Ferreira, H.J., Vizoso, M., Sayols, S., Gomez, A., Moran, S., Boque-Sastre, R., Guil, S., Martinez-Cardus, A., et al. (2016). Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. *17*.

Hilscher, C. (2005). Faster quantitative real-time PCR protocols may lose sensitivity and show increased variability. Nucleic Acids Res. *33*, e182–e182.

Holemon, H., Korshunova, Y., Ordway, J.M., Bedell, J.A., Citek, R.W., Lakey, N., Leon, J., Finney, M., McPherson, J.D., and Jeddeloh, J.A. (2007). MethylScreen: DNA methylation density monitoring using quantitative PCR. BioTechniques *43*, 683–693.

Iacobucci, I., and Mullighan, C.G. (2017). Genetic Basis of Acute Lymphoblastic Leukemia. J. Clin. Oncol. *35*, 975–983.

Inaba, H., Greaves, M., and Mullighan, C.G. (2013). Acute lymphoblastic leukaemia. The Lancet *381*, 1943–1955.

Janssen, J.W., Ludwig, W.D., Sterry, W., and Bartram, C.R. (1993). SIL-TAL1 deletion in T-cell acute lymphoblastic leukemia. Leukemia *7*, 1204–1210.

Jia, Y., Chng, W.-J., and Zhou, J. (2019). Super-enhancers: critical roles and therapeutic targets in hematologic malignancies. J. Hematol. Oncol.J Hematol Oncol *12*, 77.

Jiang, S., and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data. Brief. Funct. Genomics *17*, 104–115.

Kamachi, Y., and Kondoh, H. (2013). Sox proteins: regulators of cell fate specification and differentiation. Development *140*, 4129–4144.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 531210.

Khan, A., Mathelier, A., and Zhang, X. (2018a). Super-enhancers are transcriptionally more active and cell type-specific than stretch enhancers. Epigenetics *13*, 910–922.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018b). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. *46*, D260–D266.

Kim, H.-Y. (2014). Statistical notes for clinical researchers: Nonparametric statistical methods: 1. Nonparametric methods for comparing two groups. Restor. Dent. Endod. *39*, 235–239.

King, A.D., Huang, K., Rubbi, L., Liu, S., Wang, C.-Y., Wang, Y., Pellegrini, M., and Fan, G. (2016). Reversible Regulation of Promoter and Enhancer Histone Landscape by DNA Methylation in Mouse Embryonic Stem Cells. Cell Rep. *17*, 289–302.

Korbie, D.J., and Mattick, J.S. (2008). Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat. Protoc. *3*, 1452–1456.

Koyanagi, M., Kawakabe, S., and Arimura, Y. (2016). A comparative study of colorimetric cell proliferation assays in immune cells. Cytotechnology *68*, 1489–1498.

Kralik, P., and Ricchi, M. (2017). A Basic Guide to Real Time PCR in Microbial Diagnostics: Definitions, Parameters, and Everything. Front. Microbiol. *8*.

Kraszewska, M.D., Dawidowska, M., Larmonie, N.S.D., Kosmalska, M., Sędek, Ł., Szczepaniak, M., Grzeszczak, W., Langerak, A.W., Szczepański, T., and Witt, M. (2012). DNA methylation pattern is altered in childhood T-cell acute lymphoblastic leukemia patients as compared with normal thymic subsets: insights into CpG island methylator phenotype in T-ALL. Leukemia *26*, 367–371.

Kulis, M., and Esteller, M. (2010). 2 - DNA Methylation and Cancer. In Advances in Genetics, Z. Herceg, and T. Ushijima, eds. (Academic Press), pp. 27–56.

Kwak, S.K., and Kim, J.H. (2017). Statistical data preparation: management of missing values and outliers. Korean J. Anesthesiol. *70*, 407–411.

Lavrovsky, V.A., Guvakova, M.A., and Lavrovsky, Y.V. (1992). High frequency of tumour cell reversion to non-tumorigenic phenotype. Eur. J. Cancer *28*, 17–21.

Lee, S., and Lee, D.K. (2018). What is the proper way to apply the multiple comparison test? Korean J. Anesthesiol. *71*, 353–360.

Leggett, R.M., and Clark, M.D. (2017). A world of opportunities with nanopore sequencing. J. Exp. Bot. *68*, 5419–5429.

Leukaemia Foundation (2019). Acute Lymphoblastic Leukaemia.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Litt, M., Patel, B., Li, Y., Qiu, Y., and Huang, S. (2013). Molecular Morphogenesis of T-Cell Acute Leukemia. T-Cell Leuk. - Charact. Treat. Prev.

Litzow, M.R., and Ferrando, A.A. (2015). How I treat T-cell acute lymphoblastic leukemia in adults. Blood *126*, 833–841.

Lorenz, T.C. (2012). Polymerase Chain Reaction: Basic Protocol Plus Troubleshooting and Optimization Strategies. J. Vis. Exp. JoVE.

Luczak, M.W., and Jagodziński, P.P. (2006). The role of DNA methylation in cancer development. Folia Histochem. Cytobiol. *44*, 143–154.

Lugli, E., Roederer, M., and Cossarizza, A. (2010). Data Analysis in Flow Cytometry: The Future Just Started. Cytom. Part J. Int. Soc. Anal. Cytol. *77*, 705–713.

Mallona, I., Díez-Villanueva, A., and Peinado, M.A. (2014). Methylation plotter: a web tool for dynamic visualization of DNA methylation data. Source Code Biol. Med. *9*, 11.

Malouf, C., and Ottersbach, K. (2018). Molecular processes involved in B cell acute lymphoblastic leukaemia. Cell. Mol. Life Sci. *75*, 417–446.

Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B., et al. (2014). An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science *346*, 1373–1377.

Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Front. Genet. *10*.

Martín-Pardillos, A., Valls Chiva, Á., Bande Vargas, G., Hurtado Blanco, P., Piñeiro Cid, R., Guijarro, P.J., Hümmer, S., Bejar Serrano, E., Rodriguez-Casanova, A., Diaz-Lagares, Á., et al. (2019). The role of clonal communication and heterogeneity in breast cancer. BMC Cancer *19*, 666.

McCormack, M.P., Shields, B.J., Jackson, J.T., Nasa, C., Shi, W., Slater, N.J., Tremblay, C.S., Rabbitts, T.H., and Curtis, D.J. (2013). Requirement for Lyl1 in a model of Lmo2-driven early T-cell precursor ALL. Blood *122*, 2093–2103.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. *33*, 5868–5877.

Melnikov, A.A., Gartenhaus, R.B., Levenson, A.S., Motchoulskaia, N.A., and Levenson (Chernokhvostov), V.V. (2005). MSRE-PCR for analysis of gene-specific DNA methylation. Nucleic Acids Res. *33*, e93.

Moharram, S.A., Shah, K., and Kazi, J.U. (2017). T-cell Acute Lymphoblastic Leukemia Cells Display Activation of Different Survival Pathways. J. Cancer *8*, 4124.

Moore, L.D., Le, T., and Fan, G. (2013). DNA Methylation and Its Basic Function. Neuropsychopharmacology *38*, 23–38.

Nahm, F.S. (2016). Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean J. Anesthesiol. *69*, 8–14.

Navarrete-Meneses, M. del P., and Pérez-Vera, P. (2017). Epigenetic alterations in acute lymphoblastic leukemia. Bol. Méd. Hosp. Infant. México Engl. Ed. *74*, 243–264.

Navarro, J.-M., Touzart, A., Pradel, L.C., Loosveld, M., Koubi, M., Fenouil, R., Noir, S.L., Maqbool, M.A., Morgado, E., Gregoire, C., et al. (2015). Site- and allele-specific polycomb dysregulation in T-cell leukaemia. Nat. Commun. *6*, 1–11.

Neelakantan, D., Drasin, D.J., and Ford, H.L. (2015). Intratumoral heterogeneity: Clonal cooperation in epithelial-to-mesenchymal transition and metastasis. Cell Adhes. Migr. *9*, 265–276.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature *543*, 72–77.

Noguchi, S., Arakawa, T., Fukuda, S., Furuno, M., Hasegawa, A., Hori, F., Ishikawa-Kato, S., Kaida, K., Kaiho, A., Kanamori-Katayama, M., et al. (2017). FANTOM5 CAGE profiles of human and mouse samples. Sci. Data *4*, 170112.

Nordlund, J., Bäcklin, C.L., Wahlberg, P., Busche, S., Berglund, E.C., Eloranta, M.-L., Flaegstad, T., Forestier, E., Frost, B.-M., Harila-Saari, A., et al. (2013). Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biol. *14*, r105.

Oakes, C.C., La Salle, S., Robaire, B., and Trasler, J.M. (2006). Evaluation of a Quantitative DNA Methylation Analysis Technique using Methylation-Sensitive/Dependent Restriction Enzymes and Real-Time PCR. Epigenetics *1*, 146–152.

O'Neil, J., Shank, J., Cusson, N., Murre, C., and Kelliher, M. (2004). TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. Cancer Cell *5*, 587–596.

Orsini, P., Minervini, C.F., Cumbo, C., Anelli, L., Zagaria, A., Minervini, A., Coccaro, N., Tota, G., Casieri, P., Impera, L., et al. (2018). Design and MinION testing of a nanopore targeted gene sequencing panel for chronic lymphocytic leukemia. Sci. Rep. *8*.

Palazzo, A.F., and Gregory, T.R. (2014). The Case for Junk DNA. PLoS Genet. *10*, e1004351.

Palomero, T., Odom, D.T., O'Neil, J., Ferrando, A.A., Margolin, A., Neuberg, D.S., Winter, S.S., Larson, R.S., Li, W., Liu, X.S., et al. (2006). Transcriptional regulatory networks downstream of TAL1/SCL in T-cell acute lymphoblastic leukemia. Blood *108*, 986–992.

Pandey, R.V., Walter, P., Kallmeyer, R., Beikircher, G., Pabinger, S., Kriegner, A., and Weinhäusel, A. (2016). MSRE-HTPrimer: a high-throughput and genome-wide primer design pipeline optimized for epigenetic research. Clin. Epigenetics *8*.

Park, S.-J., Kim, J.-H., Yoon, B.-H., and Kim, S.-Y. (2017). A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages. Genomics Inform. *15*, 11–18.

Passaro, D., Quang, C.T., and Ghysdael, J. (2016). Microenvironmental cues for T-cell acute lymphoblastic leukemia development. Immunol. Rev. *271*, 156–172.

Patel, B., Kang, Y., Cui, K., Litt, M., Riberio, M.S.J., Deng, C., Salz, T., Casada, S., Fu, X., Qiu, Y., et al. (2014). Aberrant TAL1 activation is mediated by an interchromosomal interaction in human T-cell acute lymphoblastic leukemia. Leukemia *28*, 349–361.

Porcher, C., Chagraoui, H., and Kristiansen, M.S. (2017). SCL/TAL1: a multifaceted regulator from blood development to disease. Blood *129*, 2051–2060.

Qu, H., and Fang, X. (2013). A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. Genomics Proteomics Bioinformatics *11*, 135–141.

Qu, Y., Lennartsson, A., Gaidzik, V.I., Deneberg, S., Bengtzén, S., Arzenani, M.K., Bullinger, L., Döhner, K., and Lehmann, S. (2012). Genome-Wide DNA Methylation Analysis Shows Enrichment of Differential Methylation in "Open Seas" and Enhancers and Reveals Hypomethylation in DNMT3A Mutated Cytogenetically Normal AML (CN-AML). Blood *120*, 653–653.

Quah, B.J.C., and Parish, C.R. (2010). The Use of Carboxyfluorescein Diacetate Succinimidyl Ester (CFSE) to Monitor Lymphocyte Proliferation. J. Vis. Exp. JoVE.

Quah, B.J.C., Warren, H.S., and Parish, C.R. (2007). Monitoring lymphocyte proliferation in vitro and in vivo with the intracellular fluorescent dye carboxyfluorescein diacetate succinimidyl ester. Nat. Protoc. *2*, 2049–2056.

Ramaker, R.C., Lasseigne, B.N., Hardigan, A.A., Palacio, L., Gunther, D.S., Myers, R.M., and Cooper, S.J. (2017). RNA sequencing-based cell proliferation analysis across 19 cancers identifies a subset of proliferation-informative cancers with a common survival signature. Oncotarget *8*, 38668–38681.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Rao, X., Huang, X., Zhou, Z., and Lin, X. (2013). An improvement of the 2ˆ(–delta delta CT) method for quantitative real-time polymerase chain reaction data analysis. Biostat. Bioinforma. Biomath. *3*, 71–85.

Rauscher, G.H., Kresovich, J.K., Poulin, M., Yan, L., Macias, V., Mahmoud, A.M., Al-Alem, U., Kajdacsy-Balla, A., Wiley, E.L., Tonetti, D., et al. (2015). Exploring DNA methylation changes in promoter, intragenic, and intergenic regions as early and late events in breast cancer formation. BMC Cancer *15*.

Rivera-Reyes, A., Hayer, K.E., and Bassing, C.H. (2016). Genomic Alterations of Non-Coding Regions Underlie Human Cancer: Lessons from T-ALL. Trends Mol. Med. *22*, 1035–1046.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Rogers-Broadway, K.-R., and Karteris, E. (2015). Amplification efficiency and thermal stability of qPCR instrumentation: Current landscape and future perspectives. Exp. Ther. Med. *10*, 1261–1264.

Sanda, T., and Leong, W.Z. (2017). TAL1 as a master oncogenic transcription factor in T-cell acute lymphoblastic leukemia. Exp. Hematol. *53*, 7–15.

Sanda, T., Lawton, L.N., Barrasa, M.I., Fan, Z.P., Kohlhammer, H., Gutierrez, A., Ma, W., Tatarek, J., Ahn, Y., Kelliher, M.A., et al. (2012). Core Transcriptional Regulatory Circuit Controlled by the TAL1 Complex in Human T-cell Acute Lymphoblastic Leukemia. Cancer Cell *22*.

Sarantopoulou, D., Nayak, S., Brooks, T.G., Lahens, N.F., and Grant, G.R. (2019). Comparative evaluation of full-length isoform quantification from RNA-Seq. BioRxiv 698605.

Scacheri, C.A., and Scacheri, P.C. (2015). Mutations in the non-coding genome. Curr. Opin. Pediatr. *27*, 659–664.

Sharifi-Zarchi, A., Gerovska, D., Adachi, K., Totonchi, M., Pezeshk, H., Taft, R.J., Schöler, H.R., Chitsaz, H., Sadeghi, M., Baharvand, H., et al. (2017). DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC Genomics *18*.

Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J.M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Phan, T.G., Junankar, S., et al. (2019). High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat. Commun. *10*, 1–13.

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. *9*, 477–485.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*, D726–D732.

Song, Y., Berg, P.R. van den, Markoulaki, S., Soldner, F., Dall'Agnese, A., Henninger, J.E., Drotar, J., Rosenau, N., Cohen, M.A., Young, R.A., et al. (2019). Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs. Mol. Cell *75*, 905-920.e6.

Sreedharan, S.P., Kumar, A., and Giridhar, P. (2018). Primer design and amplification efficiencies are crucial for reliability of quantitative PCR studies of caffeine biosynthetic N-methyltransferases in coffee. 3 Biotech *8*.

Starmer, J., and Magnuson, T. (2016). Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. BMC Bioinformatics *17*.

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. Brief. Bioinform. *17*, 953–966.

Stockholm, D., Benchaouir, R., Picot, J., Rameau, P., Neildez, T.M.A., Landini, G., Laplace-Builhé, C., and Paldi, A. (2007). The Origin of Phenotypic Heterogeneity in a Clonal Cell Population In Vitro. PLoS ONE *2*, e394.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. *100*, 9440–9445.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. *12*.

Sun, C., Wang, L., Wang, H., Huang, T., and Zhang, X. (2019). Single-cell RNA-seq highlights heterogeneity in human primary Wharton's Jelly mesenchymal stem/stromal cells cultured in vitro. BioRxiv 723130.

Tan, T.K., Zhang, C., and Sanda, T. (2019). Oncogenic transcriptional program driven by TAL1 in T-cell acute lymphoblastic leukemia. Int. J. Hematol. *109*, 5–17.

Teif, V.B., Mallm, J.-P., Sharma, T., Mark Welch, D.B., Rippe, K., Eils, R., Langowski, J., Olins, A.L., and Olins, D.E. (2017). Nucleosome repositioning during differentiation of a human myeloid leukemia cell line. Nucleus *8*, 188–204.

Teo, Y.Y., Small, K.S., Fry, A.E., Wu, Y., Kwiatkowski, D.P., and Clark, T.G. (2009). Power consequences of linkage disequilibrium variation between populations. Genet. Epidemiol. *33*, 128–135.

Terwilliger, T., and Abdul-Hay, M. (2017). Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood Cancer J. *7*, e577–e577.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

The ENCODE Project Consortium (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. *9*, e1001046.

Theuns, S., Vanmechelen, B., Bernaert, Q., Deboutte, W., Vandenhole, M., Beller, L., Matthijnssens, J., Maes, P., and Nauwynck, H.J. (2018). Nanopore sequencing as a revolutionary diagnostic tool for porcine viral enteric disease complexes identifies porcine kobuvirus as an important enteric virus. Sci. Rep. *8*, 1–13.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Tippens, N.D., Vihervaara, A., and Lis, J.T. (2018). Enhancer transcription: what, where, when, and why? Genes Dev. *32*, 1–3.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. *19*, A68–A77.

Tran, A., Escovedo, C., Migdall-Wilson, J., Chou, A.P., Chen, W., Cloughesy, T., Nelson, S., and Lai, A. (2010). In Silico Enhanced Restriction Enzyme Based Methylation Analysis of the Human Glioblastoma Genome Using Agilent 244K CpG Island Microarrays. Front. Neurosci. *3*.

Trivedi, U.H., Cézard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., and Gharbi, K. (2014). Quality control of next-generation sequencing data without a reference. Front. Genet. *5*.

Tuğrul, M., Paixão, T., Barton, N.H., and Tkačik, G. (2015). Dynamics of Transcription Factor Binding Site Evolution. PLOS Genet. *11*, e1005639.

Van Vlierberghe, P., and Ferrando, A. (2012). The molecular basis of T cell acute lymphoblastic leukemia. J. Clin. Invest. *122*, 3398–3406.

Vandiver, A.R., Idrizi, A., Rizzardi, L., Feinberg, A.P., and Hansen, K.D. (2015). DNA methylation is stable during replication and cell cycle arrest. Sci. Rep. *5*, 17911.

VanLiere, J.M., and Rosenberg, N.A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. Theor. Popul. Biol. *74*, 130–137.

Varley, K.E., Gertz, J., Bowling, K.M., Parker, S.L., Reddy, T.E., Pauli-Behn, F., Cross, M.K., Williams, B.A., Stamatoyannopoulos, J.A., Crawford, G.E., et al. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. *23*, 555–567.

Wagner, E.M. (2013). Monitoring Gene Expression: Quantitative Real-Time RT-PCR. In Lipoproteins and Cardiovascular Disease: Methods and Protocols, L.A. Freeman, ed. (Totowa, NJ: Humana Press), pp. 19–45.

Waldman, Y.Y., Geiger, T., and Ruppin, E. (2013). A Genome-Wide Systematic Analysis Reveals Different and Predictive Proliferation Expression Signatures of Cancerous vs. Non-Cancerous Cells. PLOS Genet. *9*, e1003806.

Wall, J.D., and Pritchard, J.K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet. *4*, 587–597.

Wei, S., Weiss, Z.R., and Williams, Z. (2018). Rapid Multiplex Small DNA Sequencing on the MinION Nanopore Sequencing Platform. G3 Genes Genomes Genet. *8*, 1649–1657.

Weigel, C., Chaisaingmongkol, J., Assenov, Y., Kuhmann, C., Winkler, V., Santi, I., Bogatyrova, O., Kaucher, S., Bermejo, J.L., Leung, S.Y., et al. (2019). DNA methylation at an enhancer of the three prime repair exonuclease 2 gene (TREX2) is linked to gene expression and survival in laryngeal cancer. Clin. Epigenetics *11*, 67.

Witte, S., Bradley, A., Enright, A.J., and Muljo, S.A. (2015). High-density P300 enhancers control cell state transitions. BMC Genomics *16*.

Wu, D.-Y., Bittencourt, D., Stallcup, M.R., and Siegmund, K.D. (2015). Identifying differential transcription factor binding in ChIP-seq. Front. Genet. *6*.

Yadav, B.D., Samuels, A.L., Wells, J.E., Sutton, R., Venn, N.C., Bendak, K., Anderson, D., Marshall, G.M., Cole, C.H., Beesley, A.H., et al. (2016). Heterogeneity in mechanisms of emergent resistance in pediatric T-cell acute lymphoblastic leukemia. Oncotarget *7*.

Yamamoto, K., Nakamachi, Y., Yakushijin, K., Miyata, Y., Okamura, A., Kawano, S., Matsuoka, H., and Minami, H. (2013). A novel TRB@/NOTCH1 fusion gene in T-cell lymphoblastic lymphoma with t(7;9)(q34;q34). Eur. J. Haematol. *90*, 68–75.

Yamamura, A., Matsuo, J., Lim, Y.H.M., Mon, N.N., Heng, D.L., Unno, M., Yeoh, K.G., Osato, M., and Ito, Y. (2017). Abstract LB-138: Runx1 enhancer element marks stem cells in multiple organs. Cancer Res. *77*, LB-LB-138.

You, M.J., Medeiros, L.J., and Hsi, E.D. (2015). T-Lymphoblastic Leukemia/Lymphoma. Am. J. Clin. Pathol. *144*, 411–422.

Zacher, B., Michel, M., Schwalb, B., Cramer, P., Tresch, A., and Gagneur, J. (2017). Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. PLoS ONE *12*.

Zhang, C., Zhang, B., Lin, L.-L., and Zhao, S. (2017a). Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics *18*, 583.

Zhang, Q., Zeng, X., Younkin, S., Kawli, T., Snyder, M.P., and Keleş, S. (2016). Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. BMC Bioinformatics *17*.

Zhang, S., Wang, B., Wan, L., and Li, L.M. (2017b). Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling. BMC Bioinformatics *18*.

Zhang, Z., Murtagh, F., Van Poucke, S., Lin, S., and Lan, P. (2017c). Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. Ann. Transl. Med. *5*.

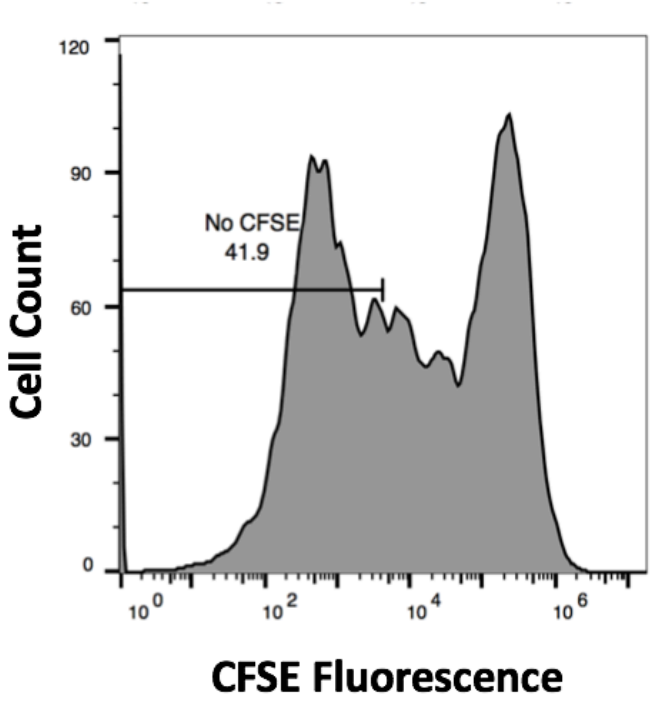
Zhao, Y., Waldman, S.D., and Flynn, L.E. (2012). The Effect of Serial Passaging on the Proliferation and Differentiation of Bovine Adipose-Derived Stem Cells. Cells Tissues Organs *195*, 414–427.

Zheng-Bradley, X., and Flicek, P. (2017). Applications of the 1000 Genomes Project resources. Brief. Funct. Genomics *16*, 163–170.

Chapter 7 - Appendix

7.1 – Chapter 2: Supplementary Figures, Tables and Command Lines

7.1.1 Supplementary Figures



| Figure 7.1. Flow cytometry histogram plots of CFSE stained Jurkat Clone 2 at cell concentration of 2x10^4/well 48-hour growth in 10% FBS 1% PSF in RPMI 1640 with 0.5µM of CFSE.
Gates were established based on the negative control (no CFSE stained sample) and display the percentage of cells within the total population that fall within the fluorescence of the negative control (41.9). Cell count is demonstrated on the x-axis and CFSE fluorescence intensity is plotted logarithmically along the y-axis.

7.1.2 Supplementary Tables

| Table 7.1. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Holm FWER method between all Jurkat cell lines at passage 1 for averaged proliferation index

Cell lines	C1	C10	C11	C2	C3	C4	C5	C6	C8	C9	P0
C10	1										
C11	1	1									
C2	1	1	1								
C3	1	1	1	1							
C4	1	1	1	1	1						
C5	1	1	1	1	1	1					

C6	1	1	1	1	1	1	1				
C8	1	1	1	1	1	1	1	1			
C9	1	1	1	1	1	1	1	1	1		
P0	1	1	1	1	1	1	1	1	0.53442	1	
P00	1	1	1	1	1	1	1	1	1	1	1

| Table 7.2. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Benjamini-Hochberg FDR method between all Jurkat cell lines at passage 1 averaged proliferation index

Cell lines	C1	C10	C11	C2	C3	C4	C5	C6	C8	C9	P0
C10	0.67262										
C11	0.83065	0.5814									
C2	0.6431	0.98725	0.5814								
C3	0.657	0.98725	0.5814	0.98725							
C4	0.98725	0.59575	0.94972	0.59372	0.59372						
C5	0.95577	0.59372	0.97894	0.59372	0.59372	0.98725					
C6	0.97894	0.59372	0.95577	0.59372	0.59372	0.98725	0.98725				
C8	0.5555	0.78102	0.40371	0.79761	0.78341	0.44764	0.42397	0.4337			
C9	0.657	0.98725	0.5814	0.98725	1	0.59372	0.59372	0.59372	0.78341		
P0	0.545 45	0.24418	0.59575	0.24418	0.24418	0.5814	0.59372	0.59372	0.09517	0.24418	
P00	0.6431	0.42286	0.84623	0.41708	0.41708	0.72328	0.77103	0.74511	0.24418	0.41708	0.79341

| Table 7.3. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Holm FWER method between all Jurkat cell lines at passage 5 averaged proliferation index

Cell lines	C1	C10	C11	C2	C3	C4	C5	C6	C8	C9	P0
C10	1										
C11	1	1									
C2	1	1	1								
C3	1	1	1	1							
C4	1	1	1	1	1						
C5	1	1	1	1	1	1					
C6	1	1	1	1	1	1	1				
C8	1	0.66594	0.30783	1	1	1	1	1			
C9	1	1	1	1	1	1	1	1	1		
P0	1	1	1	1	1	1	1	1	1	1	

P00	1	1	1	1	1	1	1	1	1	1	1

| Table 7.4. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Benjamini-Hochberg FDR method between all Jurkat cell lines at passage 5 averaged proliferation index

Cell lines	C1	C10	C11	C2	C3	C4	C5	C6	C8	C9	P0
C10	0.50383										
C11	0.41814	0.85971									
C2	0.70201	0.72089	0.64868								
C3	0.89006	0.42653	0.33137	0.64868							
C4	0.63775	0.21791	0.16053	0.36321	0.68146						
C5	0.6063	0.21791	0.16053	0.33137	0.64868	0.96015					
C6	0.89006	0.42653	0.33137	0.64868	1	0.68146	0.64868				
C8	0.41814	0.13852	0.13852	0.21791	0.46345	0.72089	0.76465	0.46345			
C9	0.80874	0.64868	0.5617	0.85971	0.72089	0.45089	0.42653	0.72089	0.27506		
P0	0.63775	0.85971	0.72089	0.85971	0.5322	0.27506	0.26528	0.5322	0.16053	0.72089	
P00	0.64868	0.23717	0.17195	0.39729	0.70201	0.96015	0.93174	0.70201	0.72089	0.46345	0.30168

| Table 7.5. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Holm FWER method between all Jurkat cell lines at passage 9 averaged proliferation index

Cell lines	C1	C10	C11	C2	C3	C4	C5	C6	C8	C9	P0
C10	1										
C11	1	1									
C2	1	1	1								
C3	1	1	1	1							
C4	1	1	1	1	1						
C5	1	1	1	1	1	1					
C6	1	1	1	1	1	1	1				
C8	1	1	1	1	1	1	1	1			
C9	1	1	1	1	1	1	1	1	1		
P0	1	1	1	1	1	1	1	1	1	0.95643	
P00	1	1	0.95643	1	1	1	1	1	1	0.45638	1

| Table 7.6. Kruskal-Wallis Post-Hoc Dunn Test, further adjusted by Benjamini-Hochberg FDR method between all Jurkat cell lines at passage 9 averaged proliferation index

Cell lines	C1	C10	C11	C2	C3	C4	C5	C6	C8	C9	P0
C10	0.83647										
C11	0.74378	0.83647									
C2	0.67463	0.5756	0.46361								
C3	0.46361	0.37266	0.2857	0.74378							

C4	0.37266	0.2857	0.21775	0.67463	0.83647						
C5	0.69432	0.60567	0.46361	0.94784	0.73509	0.66994					
C6	0.86664	0.76012	0.6911	0.73509	0.56097	0.45595	0.74378				
C8	0.6063	0.46361	0.37266	0.86664	0.83647	0.73509	0.83647	0.67463			
C9	0.67463	0.74378	0.83647	0.37266	0.21775	0.18891	0.37266	0.60567	0.31211		
P0	0.37163	0.27692	0.20387	0.65376	0.83647	0.93381	0.60779	0.39294	0.7109	0.16802	
P00	0.27889	0.20387	0.16802	0.54664	0.73509	0.83647	0.5041	0.34073	0.60779	0.15613	0.83647

| Table 7.7. List of primers sets for qPCR analysis of genes *TAL1*, *GATA3*, *MYB* and *RUNX1*

Gene	Forward Primer 5'-3'	Reverse Primer 5'-3'
TAL1	ccccctatgagatggagatt	aaaggccccgttcacatt
GATA3	ctctctgctcttcgctaccc	gcgacgactctgcaattct
MYB	tggaccaaagaagaagatcaga	tctccccttaagtgcttgg
RUNX1	actcggctgagctgagaaatg	gacttgcggtgggtttgtg

7.1.2.a qPCR primer sets and amplicon sequences (Methods: 2.2.9)

TAL1

Exon 4-5 (Intron 4)

841 ccccctatga gatggagatt actgatggtc cccacaccaa agttgtgcgg cgtatcttca
901 ccaacagccg ggagcgatgg cggcagcaga atgtgaacgg ggcctttgcc gagctccgca

>chr1:47685767-47689702 3936bp CCCCCTATGAGATGGAGATT
AAAGGCCCCGTTCACATT
CCCCCTATGAGATGGAGATTactgatggtgagtctgccccaccccaccgt
tgtacctgcccctcacgctggttagagccctctggaggagggaacatctg
ggactactccgtaggtatctcacaaagatgggcgggccactttccaaacc
caaactgccgaagcccttcaccctgcagctgagctagggctaagtaaagt
cctggatggggctagcaggaccacactcacagtcctgcaaagaggaggca
tgggagagagacacataatggcgcttagcccaaacctctggagttagttc
ccaaacctttggagttagttccagtccccagtccgggctccattcctaca
ctttggcaaggatcccaagtgggcgcagaatcacacagtgtcccccgatc
ccccattctaggcactgaccctcagggattccctgcacttcctcccaagt
cttccttagagtctggagtctgcttcctccctgcaatggagtcatcaagg
gttttctccgaggtccctacacagtaggctcccccttgtctcccgttaaca
cattctgggcttagacctggagttcccttcctgacatcatggattctacc
ttgccccctggattccccgagaagggccttggcactgctcctttgaggta
cttgaatattccctgaacttgcaggggtggtttggacaaatgggagattt
ggaagaccctcgccatctttccaaggttgaccttgtatagtagattggtc
ccagcaccaggggtaggagagcaggcacttcagagctgagggactttgag
ccagacccagatgcccccagaacttcagcttggctgtggcctggggctcc
tgtaggaaggctgcccaaggcctcagagcaaggctgggatttatgctcca
aactggtgggaggaatatgagtatggagctggagcctctggagtggagcc

ctctgtggtgttctggagaggagcggggctgggatctgcacaacagaaaa
gacttcaaaagacccaagcagcctggggcacaggtgtgcagggctgagag
gcaggggtgccttctgctagattagagaggagttctctagtctaagagag
gtgggtgtaccaaacaggaagaagacaggaccagccctgaagttccccag
ggatgaggcattttctacaataggcagaagaggcagagcataggggcgag
ggtttgtggggtgtgcaggagtgtagggcagtcattatccataagtgttg
gcaatatttaccgagcacctactcagtgctggccagagagaggcagagga
caccaggaggggtagacagactcaggaaagaggaaggtgaggagaggggg
tggttctctggaataaaaccagagaggaataggcagacatgtaggcgaag
ggtagggtatgagtgtgcaggatgtgtctaggggacactgaaattagaga
ttgaacccaaaaagagtccatagaagtagaactaatgcaagatctagagt
taggacttgatgtggattcctgttgctgaattgctactaccatgtattgg
acacctactatgtgccatggactatgtgtgcttgatttctgtttattcct
taccacaagctcatgagggcaatattattgtttacatctcacagataagg
aaattgaggctgagagaggttacccagcttgatcaaggtcacacagctaa
gtcctgggtcacacagctaagtcctggaccaccctcataacactggtctt
aagctttggaataagtcctgaccacttccccaccatctccccttctggat
gtagtctagctccagctgtgacagcagtcataagtacccccctccccaact
aggagctccctcaagatcaccagcattagccatggtggagggactgtcct
ccctatgtcactgctgagggatgctctctgtctggcctggaggcctgaaa
gcagcaagtaatttcctcagaggccgcccagggtagaactgccctgcagt
tgtgctgtgtttttcccaggctgggtctggcctttggcaaacagtcagct
gggctagggcctgctatggtttggaaaggcagcttcagaggtaccatggg
caggggatggtgatccttcccattaaaagtgtgtgtgtgtgaacacccac
acatgcatgtgcatgcctgtgtgcgtgtaggagagagaaagagatgatca
ggcagtaagtgtggcatggatgagctgattcacagactgaccactgaggt
gaggtctgcttccatcccagcctcagtctccttagctataaaatgggcag
tgggacttaggtgaggtcttagctcttctcctctacctcggttttcacac
tctgaggtctggttccaggggttactagaagagggaaaataaatatccta
ggactggcagttccttcagtgctcagagcttcctggactatgtgaccctg
gcttgtgcctcgatgttagggagagaagcagtctcctcagcccagaagtt
caagagcactagtaggacccttcactgtcttcccctggctttacaaatcc
ccagtgagtcttgaacacagcacagtgctttgcacaccataagtgcccta
gacatgtgtgctggtatctgccactgctccataagtctttatacatggaa
tgttagtgactattgggccaagccatagtcagtggattgtagacatgagg
agactgaggcagccttcagaagcctgcctagagctataggccatggctgt
gcatagctgtgtcttcacttctctcctgggtatgagagcccttctgcatc
gatctctgctctcccctccattccccacctgccaccatgatgggtcagtt
agcagacattttctaagcccctactctgtgctttccatcacccttaggat
aaagtcccccccgccttactgcaaggccctgtggagtctggaccctgccta
ctctgatcttgtgcacaggactccagccagcctggcttcaggccttgtct
agtgcagggcctgtgcattcccttggcttggatactcttcttcctctccc
cttcttgtgcctggctgtttcttcacagcctcagacaagccttcccagcc
actggagctaaacatggagcttcttttcattacaattcagccctggtttg
tttgtttttaatattcgtttccccattaacatgtaagctccatgagggc
aggcactcaatttagttcaccattattcctcagtgcctagcacagtgcct
ggttgttagcaaaaaactgatacatggataataatagctaacatttattg

actgctaagctgaggcaggcattgtcttaagttctttacctgtgttaggt
actagacccatgtaacaactcatgacttagtagtattacctccattttgc
aggtgatgaagccaaggcacttcacaagtttttatgtcacttgcccagag
tcattcatttaggtaggggtagtcaaataggaatggagtctgtctcctta
gacatagctgtaaggcctcttggcacaagtgaatggatgaatgaataaca
tacggtccctgtcctcaaatatatcacagcccagggagggagacaagcac
agaaacataagtaatgatcaagagtgatggaactccagggattgcggtat
cgtggtggttggttacatctgtctggaggggttgaggaagttcccataga
tgaggaagcacaggaacttgaatttgaagggtggatgttttctaagtaaa
aggaaagtaggcattctaggcttcccaagtttcccaatgtatggagatgc
ctttggggaaggaatctcaagcccattctcctaactcttgtcctcctaat
ctccaggtccccacaccaaagttgtgcggcgtatcttcaccaacagccgg
gagcgatggcggcagcagAATGTGAACGGGGCCTTT

RUNX1

Primers taken from: Challen, G.A, and Goodell, M.A. (2010). Runx1 isoforms show differential expression patterns during hematopoietic development but have similar functional effects in adult hematopoietic stem cells. Exp Hematol *38*, 403-416.

GATA3

Exon 1 1…204 → 191…291
ADDITIONAL OLIGOS

	start	len	tm	gc%	any	3'	seq
1 LEFT PRIMER	3	20	59.33	60.00	2.00	0.00	**ctctctgctcttcgctaccc**
RIGHT PRIMER	108	19	59.13	52.63	4.00	2.00	**gcgacgactctgcaattct**

PRODUCT SIZE: 106, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 0.00

>chr10:8096833+8097333 501bp CTCTCTGCTCTTCGCTACCC GCGACGACTCTGCAATTCT
CTCTCTGCTCTTCGCTACCCaggttggtactggtgacttttttttttttt
aagtttgatttttttgcccccaaccacttgggaggacctaaatcaatttta
aaaactcaactctcctcttttggaggttttctaggggctgagaggacggt
cccgggaccggtgtccccgagggagggacttgccctccaagtcgtaacag
tcagccctgggacttgccctccaagttgctcagccagccccggctcccgc
gagccgggctgcagggacgtccccgagagccctgcgggctccgcggccgt
gtccccgcgctcccgtgcgggtctcgggtgcgctgggcgggcgggcggcg
cgaggggaggttgtgccactccagcaactcaggggctcatccaggtctcc
cattctctcccttgcaggtgacccgaggagggactccgcctccgagcggc
tgaggaccccggtgcagaggagcctggctcgcAGAATTGCAGAGTCGTCG
C

MYB

Exon 4: 413…505 → 481…591

OLIGO start len tm gc% any 3' seq
LEFT PRIMER 2 22 58.91 40.91 5.00 3.00 **tggaccaaagaagaagatcaga**
RIGHT PRIMER 94 20 59.31 50.00 5.00 0.00 **tctccccttaagtgcttgg**
SEQUENCE SIZE: 110
INCLUDED REGION SIZE: 110

PRODUCT SIZE: 93, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00

1 ttggaccaaagaagaagatcagagagtgatagagcttgtacagaaatacggtccgaaacg
>>>>>>>>>>>>>>>>>>>>

61 ttggtctgttattgccaagcacttaaaggggagaattggaaaacaatgta
<<<<<<<<<<<<<<<<<<<

>chr6:135510998+135511333 336bp TGGACCAAAGAAGAAGATCAGA
TCTCCCCTTTAAGTGCTTGG
TGGACCAAAGAAGAAGATCAGAgagtaagttctttcttcattggtgtgtg
actcataattaagaatattcccaaaatgctaatattttcctatttgagga
agcaggtaaaatgggcaaacaggaagtagaggactctattcccatatttc
cagtgaatgaaagcaaattttggaaattttctaaagatcttgtaacactg
aagaatgattatactgactcattacataactttaaaacataggttatttt
tgtgtgtttatctgaaggtgatagagcttgtacagaaatacggtccgaaa
cgttggtctgttattgCCAAGCACTTAAAGGGGAGA

| Table 7.8. Friedman Test results for averaged gene expression of TAL1 CRC genes at passages 1, 5 and 9 (p<0.05).

Genes	**TAL1**	**GATA3**	**RUNX1**	**MYB**
p-value (p<0.05)	0.01685	0.00525	0.00483	0.2778
Significance	Yes	Yes	Yes	Yes

| Table 7.9. Wilcoxon Rank Sum Two-Tailed Test as Post-Hoc for Friedman test (Table 7.8) between passages 1 and 5, 1 and 9 and 1 and 5 for averaged TAL1 CRC gene expression

TAL1 CRC Genes				
p-value (p<0.05)	TAL1	GATA3	RUNX1	MYB
P1 vs P5	0.0151	0.0151	0.0048*	0.012*
P1 vs P9	0.18	0.88	0.04*	0.238
P1 vs P5	0.006*	0.008*	0.24	0.05

[1]Significant p-values indicated by asterisk (*)

| Table 7.10. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Holm FWER method for *TAL1* gene expression between gene expression groups

Gene Expression Groups	GATA3+	RUNX1+ MYB+
RUNX1+ MYB+	0.625854	
TAL1+	0.057444	0.057444

| Table 7.11. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Benjamini-Hochberg FDR method for *TAL1* gene expression between gene expression groups

Gene Expression Groups	GATA3+	RUNX1+ MYB+
RUNX1+ MYB+	0.625854	
TAL1+	0.029888	0.029888

| Table 7.12. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Holm FWER method for *GATA3* gene expression between gene expression groups

Gene Expression Groups	GATA3+	RUNX1+ MYB+
RUNX1+ MYB+	0.018685	
TAL1+	0.127682	0.390284

| Table 7.13. Kruskal-Wallis Dunn Post-Hoc test p-values further adjusted by the Benjamini-Hochberg FDR method for *GATA3* gene expression between gene expression groups

Gene Expression Groups	GATA3+	RUNX1+ MYB+
RUNX1+ MYB+	0.018685	
TAL1+	0.095761	0.390284

| Table 7.14. Spearman's Rho Correlation test p-values for the $GATA3^{+}$ group and gene expression of TAL1 CRC genes

$GATA3^{+}$ Group	MYB	TAL1	RUNX1
GATA3	0.00026	0.00385	$3e^{-05}$
RUNX1	$3e^{-05}$	0.02082	
TAL1	0.0149		

| Table 7.15. Spearman's Rho Correlation test p-values for the $TAL1^{+}$ group and gene expression of TAL1 CRC genes

$TAL1^{+}$ Group	MYB	TAL1	RUNX1
GATA3	0.0053	0.0004	0.001
RUNX1	0.004	0.0003	
TAL1	0.00061		

| Table 7.16. Spearman's Rho Correlation test p-values for the RUNX1/MYB^{+} group and gene expression of TAL1 CRC genes

RUNX1/MYB^{+} Group	MYB	TAL1	RUNX1
GATA3	0.00014	0.00087	0.00381
RUNX1	0.00247	0.0023	
TAL1	$6e^{-05}$		

| Table 7.17. Spearman's Rho Correlation test p-values for all Jurkat parental and clonal cell lines and gene expression of TAL1 CRC genes

All Cell lines	MYB	TAL1	RUNX1
GATA3	$5.699e^{-05}$	0.0003	$1.35e^{-6}$
RUNX1	$6.58e^{-08}$	$8.84e^{-7}$	
TAL1	$3.2e^{-07}$		

| Table 7.18. Spearman's Rho Correlation test p-values for proliferation index (PI) correlation to TAL1 CRC gene expression over all passages, passage 1, 5 and 9 for all Jurkat parental and clonal cell lines

PI Correlation	All Passages	Passage 1	Passage 5	Passage 9
TAL1	0.47	0.30	0.45	0.61
GATA3	0.27	0.94	0.65	0.86
RUNX1	0.74	0.46	0.65	0.85
MYB	0.82	0.51	0.94	0.81

| Table 7.19. Spearman's Rho Correlation test p-values for proliferation index (PI) correlation to TAL1 CRC gene expression for all passages for gene expression groups

PI Correlation	GATA3+	TAL1+	RUNX1/MYB+
TAL1	0.14	0.2	0.83
GATA3	0.14	0.2	0.51
RUNX1	0.23	0.49	0.54
MYB	0.28	0.36	0.37

7.1.2.b Pheatmap Command Line – R programming language

```
library(pheatmap)

# Create test matrix
T <- data.matrix(datafile., rownames=1)
TALL <- data.matrix(datafile)

colnames(TALL) = colnames(TALL)
rownames(TALL) = rownames(TALL)
```

```
#Gene Signature, euclidean distance
pheatmap(TALL, scale = "row", clustering_distance_rows = "euclidean",
clustering_distance_cols = "euclidean")

#Gene Score, no clustering, raw data
pheatmap(TALL, cluster_row = FALSE, cluster_cols = FALSE)
```

7.1.2.c GSALightning Command Line – R Programming language

```
#GSALightning: Ultra-fast Permutation-based Gene Set Analysis
library(devtools)
# if not yet installed run: install_github("billyhw/GSALightning")
library(GSALightning)

#example data
data(expression)
data(sampleInfo)
data(targetGenes)


#write targetgenes as a list with unique elements for individual
#genes or groups of genes

l <- list(TAL1 = c('TAL1'),         GATA3 = c('GATA3'),    RUNX1 = c('RUNX1'),
        MYB = c('MYB'))

#input your data
#Use unaveraged data with unique ID names for each replicate
sampleInfo <- ALLGSA
targetGenes <- l
expression <- as.matrix(ALLPGSAExpression, rownames.force = 1,
colnames(ALLPGSAExpression))

#remove genes with 0 variance
expression <-  expression[apply(expression,1,sd) != 0,]

#run permutations
#try all three methods (methods = 'maxmean' OR 'mean' OR 'absmean')
#set "nperm = X" by calculating n/0.05 x 2, where n = the number of geneSets in your data.
#set "minsize = x, maxsize = x" to the minimum and maximum number of genes in your
geneSets.
GSALightmax <- GSALight(eset = expression, fac = factor(sampleInfo$TN), gs = targetGenes,
nperm = 160, method = 'maxmean', restandardize = FALSE, minsize = 1, maxsize = 6,
rmGSGenes = 'gene', verbose = FALSE)
```

```
#view statistical results
head(GSALight)

#Investigate distribution of results
hist(GSALightmax[,'p-value:up-regulated in Control'], main=NULL, xlab='p-value')

#restandardise results
GSALightResultsReStand <- GSALight(eset = expression, fac = factor(sampleInfo$TN), gs =
targetGenes,
                          nperm = 160, method = 'maxmean', restandardize = TRUE, minsize =
1,
                          maxsize = 6, rmGSGenes = 'gene', verbose = FALSE)

#Investigate distribution of restandardised results
hist(GSALightResultsReStand[,'p-value:up-regulated in Control'], main=NULL, xlab='p-value')

#Export all results as .csv files
write.csv(GSALightmax, file = "ALL P MaxMean Permutation Statistics.csv")
write.csv(GSALightMean, file = "Mean Permutation Statistics.csv")
write.csv(GSALight, file = "ALLP AbsMean Permutation Statistics.csv")
write.csv(GSALightResultsReStandMaxMean, file = "Restandardised MaxMean Permutation
Statistics.csv")
write.csv(GSALightResultsReStandMean, file = "Restandardised Mean Permutation
Statistics.csv")
write.csv(GSALightResultsReStand, file = "ALL P Restandardised AbsMean Permutation
Statistics.csv")
```

7.2 – Chapter 3: Supplementary Figures, Tables and Command Lines

7.2.1 Supplementary Figures

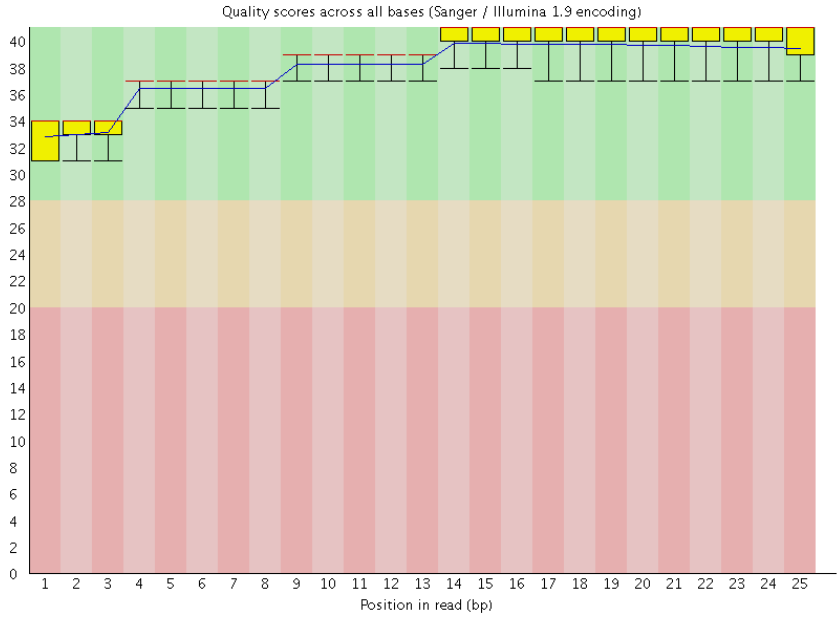
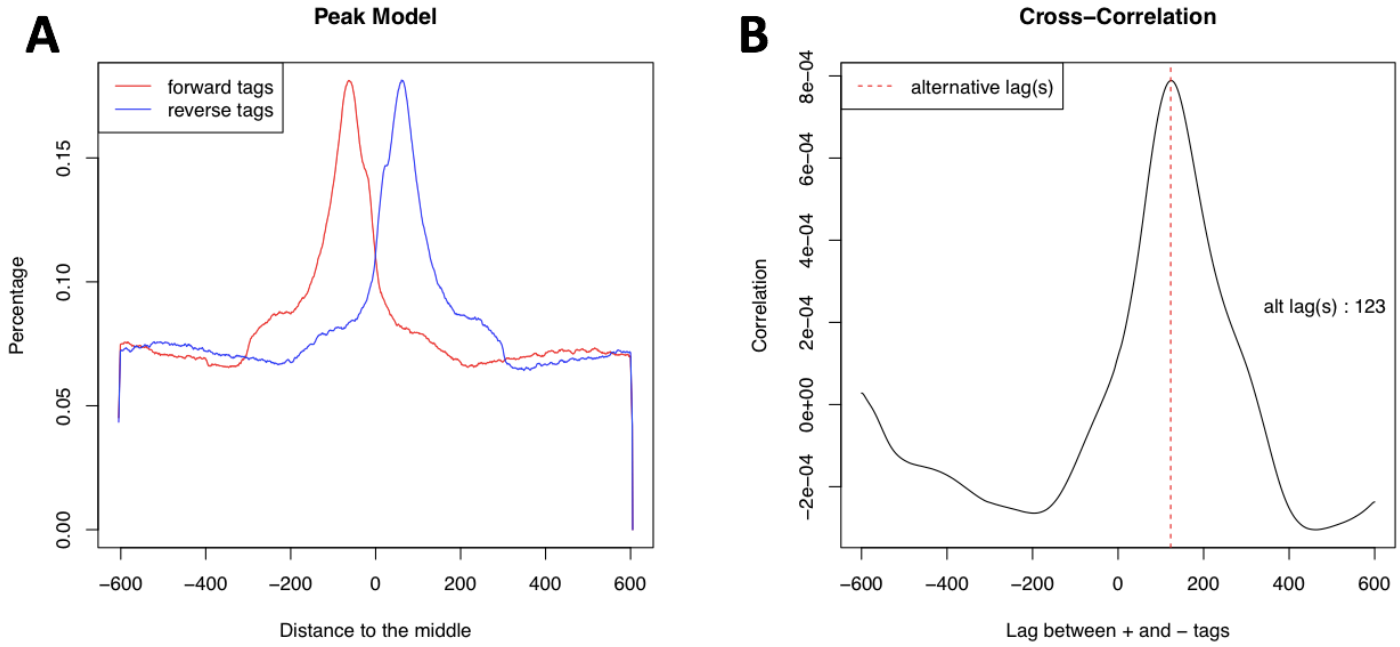
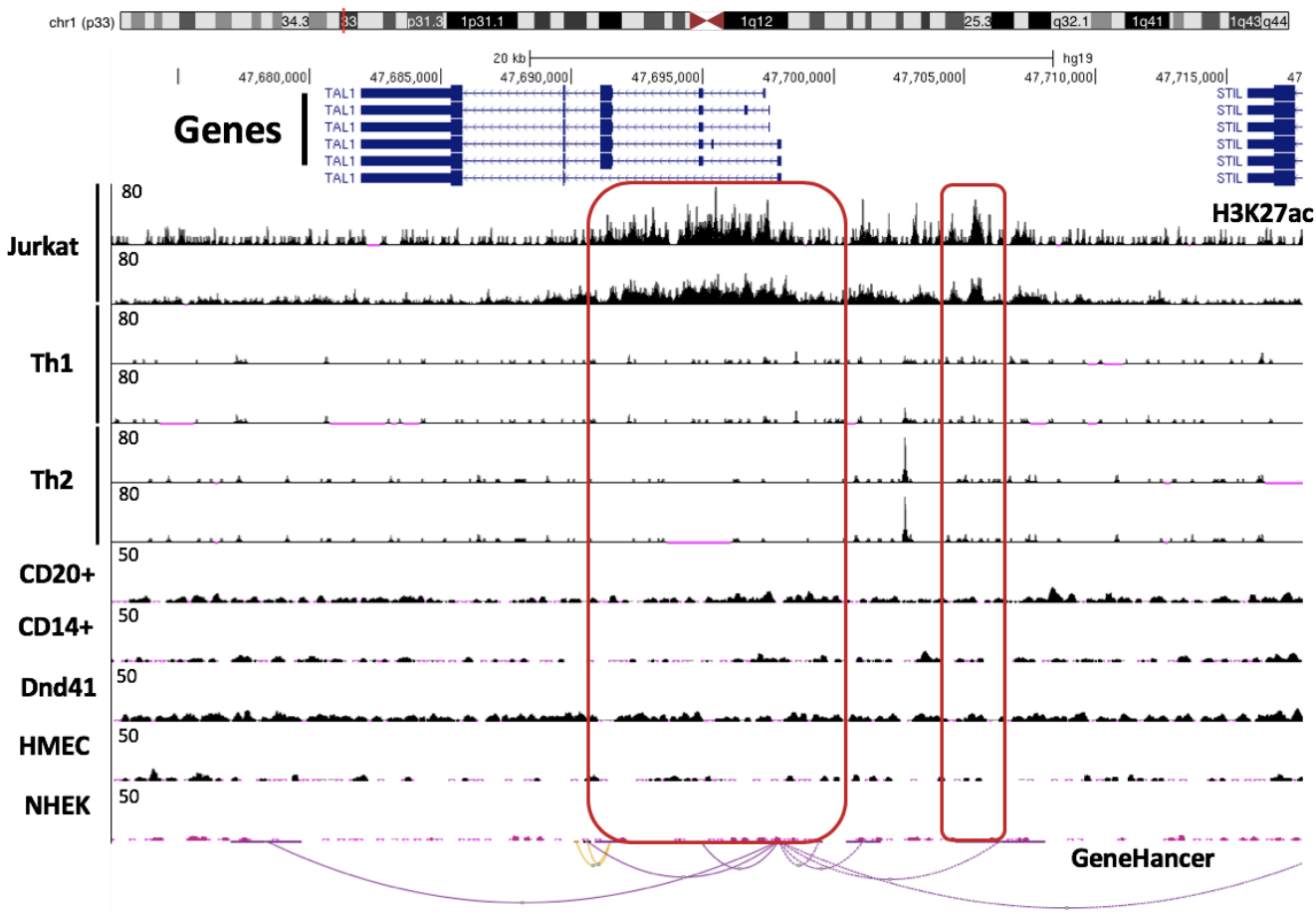


Figure 7.2. Example of the quality score across all bases for Illumina 1.9 sequencing of Jurkat H3K27ac replicate 1 (GEO Accession: GSM1431908) using the FastQC program (Method:3.2.2). Position in read per base pair (bp) along the x-axis and Phred score along y-axis. Green region indicates satisfactory quality (> 25) and orange and red regions indicate low quality of the sequencing experiment (< 25 and <20). Median is indicated by red line. Quality reads for all ChIP-seq files downloaded and tested are in Supplementary Table 7.3.2.

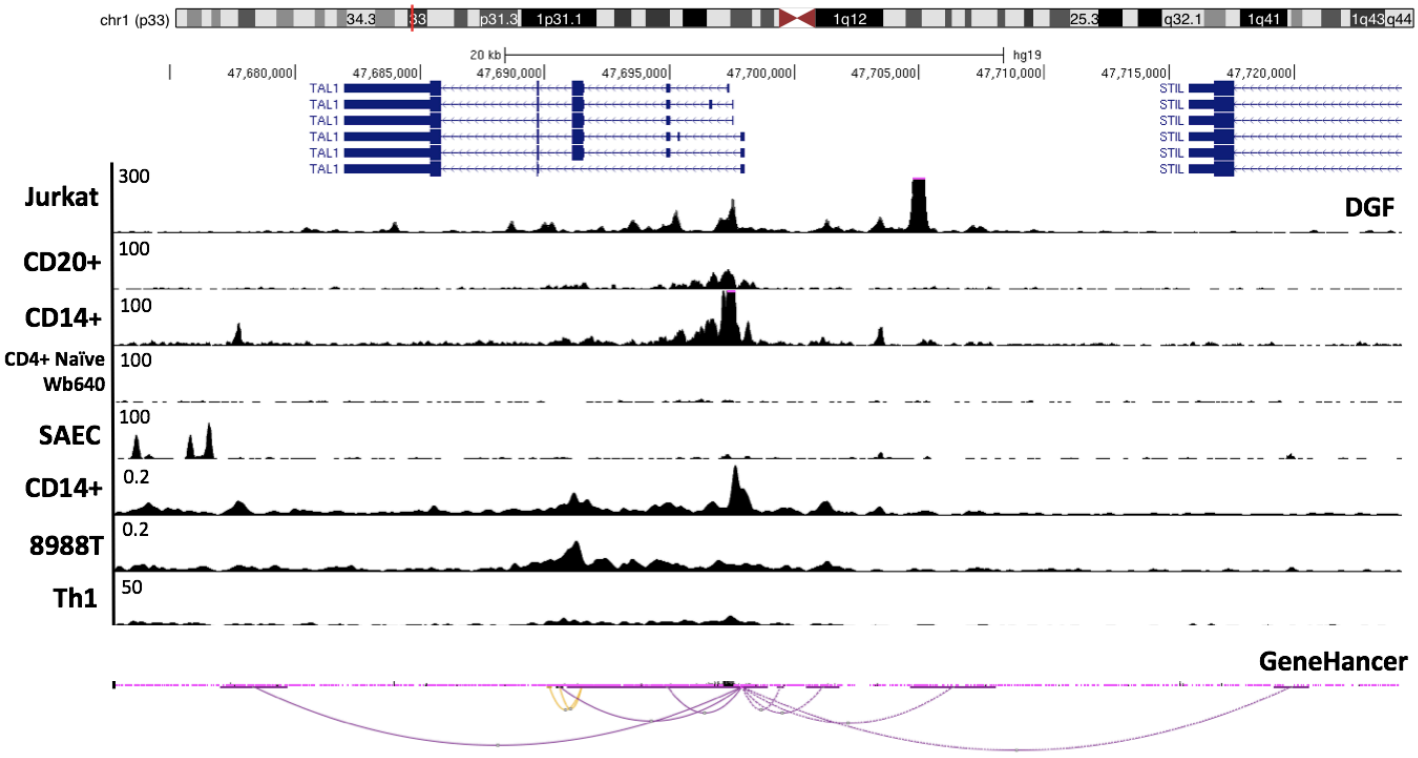


| Figure 7.3. Example of MACS2 outputs of Illumina 1.9 ChIP-seq data peak model analysis and cross-correlation analysis from the Jurkat H3K27ac replicate 1 data (GEO Accession: GSM1431908).
A. Display of peak model analysis of ChIP-seq data quality through the positional information and reads from both strands, ensuring accurately mapped locations of peaks within the data. Percentage of reads for both forward and reverse tags are displayed on the y-axis and distance of tags from the centre of a peak is shown on the x-axis. **B.** A cross-correlation plot displaying a single peak for fragment-lengths relative to read-length, ensuring correlation of tags to the sequencing reads. Y-axis displays correlation of tags and x-axis displays lag between the forward and reverse tags. List of data used and run through MACS2 can be found in Supplementary Table 7.3.1. MACS2 peak model plot and cross correlation analysis was seen to be similar for all files tested.



| Figure 7.4. H3K27ac read depth across the TAL1 locus for all replicates for cell lines, Jurkat, DND41, HMEC and NHEK, and primary cell lines Th1, Th2, CD20+ B-cells and MonoCD14+.

Chromosome ideogram of chromosome 1 indicates location of the TAL1 locus indicated by the red line. Ref-Seq genes from the ENCODE Project display isoforms of the TAL1 and STIL gene with localised ChIP-seq peaks of the cell lines listed above. Y-axis displays the read depth of H3K27ac enrichment and is normalised amongst samples from the same experiment, with primary cells and cell line samples indicated on the left-hand side (see also Appendix for details and visualisation of all replicates). GeneHancer database looping displays predicted interactions between regulatory elements across the TAL1 locus (Methods: 2.3.5). Red boxes highlight areas of H3K27ac and H3K4me3 enrichment of the Jurkat cell line relative to the other cell lines displayed.



| Figure 7.5. DNAse1 hypersensitivity (DHS) read depth displayed in the UCSC genome browser for the all replicates of Jurkat, CD20R017794, MonoCD14, CD4⁺ Naïve Wb11970640, SAEC, 8988T and Th1 cell lines across the *TAL1* locus.

Chromosome ideogram of chromosome 1 indicates location of the TAL1 locus indicates location of the TAL1 locus indicated by the red line. Ref-Seq genes from the ENCODE Project display isoforms of the TAL1 and STIL gene with localised ChIP-seq peaks of the cell lines listed above. Y-axis displays the read depth of DHS enrichment and is normalised amongst samples from the same experiment (see also Appendix for details and visualisation of all replicates). GeneHancer database looping displays relationships of regulatory elements across the TAL1 locus (Methods: 2.3.5).

7.2.2 Supplementary Tables

| Table 7.20. List of DNAse1 Hypersensitivity files downloaded from the ENCODE project used for differential binding analysis

DNAse1 Hypersensitivity						
Cell Line	**Tissue**	**File Type**	**Replicate**	**ENCODE?**	**Lab**	**GEO Accession**
8988t	Pancreas adenocarcinoma	.bam	1	Y	Duke University	GSM816667
8988t	Pancreas adenocarcinoma	.bam	2	Y	Duke University	GSM816667
8988t	Pancreas adenocarcinoma	.narrowPeak	Combined	Y	Duke University	GSM816667
Adult CD4 Th0	CD4 Th0	.bam	2	Y	Duke University	GSM1008572
Adult CD4 Th0	CD4 Th0	.bam	1	Y	Duke University	GSM1008572
Adult CD4 Th0	CD4 Th0	.narrowPeak	Combined	Y	Duke University	GSM1008572
CD20RO01794	B Cells	.bam	2	Y	Duke University	GSM1008588
CD20RO01794	B Cells	.bam	1	Y	Duke University	GSM1008588
CD20RO01794	B Cells	.narrowPeak	Combined	Y	Duke University	GSM1008588
Jurkat	T-ALL	.bam	1	Y	University of Washington	GSM736501
Jurkat	T-ALL	.bam	2	Y	University of Washington	GSM736501
Jurkat	T-ALL	.broadPeak	1	Y	University of Washington	GSM736501
Jurkat	T-ALL	.broadPeak	2	Y	University of Washington	GSM736501
MonoCD14	Monocyte	.bam	1	Y	Duke University	GSM1008582
MonoCD14	Monocyte	.bam	2	Y	Duke University	GSM1008582
MonoCD14	Monocyte	.narrowPeak	Combined	Y	Duke University	GSM1008582
SAEC	Normal Lung Epithelial	.bam	1	Y	University of Washington	GSM736617
SAEC	Normal Lung Epithelial	.bam	2	Y	University of Washington	GSM736617
SAEC	Normal Lung Epithelial	.narrowPeak	1	Y	University of Washington	GSM736617
SAEC	Normal Lung Epithelial	.narrowPeak	2	Y	University of Washington	GSM736617
Th1 v2	Th1	.bam	1	Y	University of Washington	GSM646569

Th1 v2	Th1	.bam	2	Y	University of Washington	GSM646569
Th1 v2	Th1	.broadPeak	1	Y	University of Washington	GSM646569
Th1 v2	Th1	.broadPeak	2	Y	University of Washington	GSM646569

| Table 7.21. List of H3K27ac ChIP-seq data from the ENCODE Project and other studies (GEO Accession and SRA accession)

H3K27ac ChIP-Seq						
Cell Line	**Tissue**	**File Type**	**Replicate**	**ENCODE?**	**Lab**	**GEO/SRA Accession**
CD20RO01794	B Cells	.bam	1	Y	Broad	GSM1003459
CD20RO01794	B Cells	.bam	2	Y	Broad	GSM1003459
CD20RO01794	B Cells	.broadPeak	Combined	Y	Broad	GSM1003459
DND41	T-ALL	.bam	1	Y	Broad	GSM1003462
DND41	T-ALL	.bam	2	Y	Broad	GSM1003462
DND41	T-ALL	.broadPeak	Combined	Y	Broad	GSM1003462
HMEC	Dermal Endothelim	.bam	1	Y	Broad	GSM733660
HMEC	Dermal Endothelim	.bam	2	Y	Broad	GSM733660
HMEC	Dermal Endothelim	.broadPeak	Combined	Y	Broad	GSM733660
NHEK	Keratinocytes	.bam	1	Y	Broad	GSM733674
NHEK	Keratinocytes	.bam	2	Y	Broad	GSM733674
NHEK	Keratinocytes	.bam	3	Y	Broad	GSM733674
NHEK	Keratinocytes	.broadPeak	Combined	Y	Broad	GSM733674
Jurkat	T-ALL	.fastq	1	N	HVB, Genetics and Genomics Sciences	**GSM1431908**
Jurkat	T-ALL	.fastq	2	N	HVB, Genetics and Genomics Sciences	**GSM1431909**
Th2	Th2	.fastq	1	N	Aalto University, Finland	SRR873431
Th2	Th2	.fastq	2	N	Aalto University, Finland	SRR873435
Th1	Th1	.fastq	1	N	Aalto University, Finland	SRR873432

Th1	Th1	.fastq	2	N	Aalto University, Finland	SRR873434

| Table 7.22. List of H3K4me3 ChIP-seq data from the ENCODE Project and other studies (GEO Accession and SRA Accession)

H3K4me3 ChIP-Seq						
Cell Line	**Tissue**	**File Type**	**Replicate**	**ENCODE?**	**Lab**	**GEO/SRA Accession**
Jurkat	T-ALL	.bam	1	Y	University of Washington	GSM945267
Jurkat	T-ALL	.bam	2	Y	University of Washington	GSM945267
Jurkat	T-ALL	.broadPeak	Combined	Y	University of Washington	GSM945267
SAEC	Normal Lung	.bam	1	Y	University of Washington	GSM945199
SAEC	Normal Lung	.bam	2	Y	University of Washington	GSM945199
SAEC	Normal Lung	.broadPeak	1	Y	University of Washington	GSM945199
SAEC	Normal Lung	.broadPeak	2	Y	University of Washington	GSM945199
CD20RO01794	B Cells	.bam	3	Y	University of Washington	GSM945198
CD20RO01778	B Cells	.bam	1	Y	University of Washington	GSM945229
CD20RO01778	B Cells	.broadPeak	2	Y	University of Washington	GSM945229
CD20RO01778	B Cells	.broadPeak	3	Y	University of Washington	GSM945229
MCF7	Mammary Gland	.bam	1	Y	University of Washington	GSM945269
MCF7	Mammary Gland	.bam	2	**Y**	University of Washington	GSM945269
MCF7	Mammary Gland	.broadPeak	1	**Y**	University of Washington	GSM945269
MCF7	Mammary Gland	.broadPeak	2	**Y**	University of Washington	GSM945269
Dnd41	T-ALL	.bam	1	**Y**	University of Washington	GSM1003468
Dnd41	T-ALL	.bam	2	**Y**	University of Washington	GSM1003468
Dnd41	T-ALL	.broadPeak	Combined	**Y**	University of Washington	GSM1003468

Th1	Th1	.fastq	1	Y	Aalto University, Finland	SRR873613
Th2	Th2	.fastq	1	Y	Aalto University, Finland	SRR873615

| Table 7.23. List of Input Files for ENCODE and other studies ChIP-seq data (GEO Accession and SRA Accession)

Input Files for ChIP-seq						
Cell Line	**Tissue**	**File Type**	**Replicate**	**ENCODE?**	**Lab**	**GEO/SRA Accession**
Dnd41	T-ALL	.bam	1	Y	Broad	GSM1003558
Dnd41	T-ALL	.bam	2	Y	Broad	
HMEC	Dermal Endothelim	.bam	1	Y	Broad	GSM733668
HMEC	Dermal Endothelim	.bam	2	Y	Broad	GSM733668
NHEK	Keratinocytes	.bam	1	Y	Broad	GSM733740
NHEK	Keratinocytes	.bam	2	Y	Broad	GSM733740
CD20RO01778	B Cell	.bam	1	Y	University of Washington	GSM945197
Jurkat v2	T-ALL	.fastq	1	N	HVB, Genetics and Genomics Sciences	SRR1509752
Th1	Th1	.fastq	1	Y	Aalto University, Finland	SRR873613
Th2	Th2	.fastq	1	v	Aalto University, Finland	SRR873615
Jurkat	T-ALL	.bam	1	Y	University of Washington	GSM945268
SAEC	Normal Lung	.bam	**1**	**Y**	University of Washington	GSM945314
CD20RO01794	B Cells	.bam	3	**Y**	University of Washington	GSM945195
MCF7	Mammary Gland	.bam	1	**Y**	University of Washington	GSM945274

| Table 7.24. Phred Scores of ChIP-seq data for the cell line, Jurkat and primary cell lines, Th0, Th1 and Th2 for H3K27ac, H3K4me3 and Inputs from FastQC and after FastP trimming.

Cell Type/Line	Factor	Replicate	FastQC	FastP
Jurkat	H3K27ac	1	>32	-
Jurkat	H3K27ac	2	>32	-
Jurkat	Input	1	>38	-
Th0	H3K4me3	1	21-24	22-25
Th0	Input	1	>26	-
Th0	H3K27ac	1	>32	-
Th1	H3K4me3	1	17-23	25

Th1	H3K27ac	1	>30	-
Th1	H3K27ac	2	>30	-
Th1	Input	1	>30	-
Th1	Input	2	>30	-
Th2	H3K4me3	1	19-25	25
Th2	H3K27ac	1	>30	-
Th2	H3K27ac	2	>30	-
Th2	Input	1	>30	-
Th2	Input	2	>30	-

[1]Primary cell lines and the Jurkat cell lines used are derived from Table () and tested using the FastQC program and Fastp (Supplementary Code: 7.3.3.a).

| Table 7.25. Example of the top 50 results from DiffBind Differential Binding Analysis Granges Object for primary T-cells (Th0 and Th1) and the Jurkat cell line for DNAse1 Hypersensitivity (Cell lines from Table 7.20-23).

	seqnames	start	end	width	Conc	Conc_T.Cell	Conc_T.ALL	Fold	p.value	FDR
1	chr2	5800122	5802122	2001	6.21	2.72	9.42	-6.7	1.51E-54	4.86E-49
2	chr15	25008707	25010707	2001	5.99	2.77	9.16	-6.4	5.86E-45	9.43E-40
3	chr2	59640062	59642062	2001	5	2.48	8.08	-5.6	2.35E-42	2.52E-37
4	chr8	130179473	130181473	2001	6.72	3.3	9.91	-6.61	3.44E-41	2.77E-36
5	chr7	133716225	133718225	2001	5.65	3.6	8.62	-5.02	7.76E-40	5.00E-35
6	chr8	68879953	68881953	2001	6.45	3.75	9.56	-5.82	1.78E-38	9.56E-34
7	chr8	130191098	130193098	2001	5.85	3.33	8.93	-5.6	3.07E-37	1.41E-32
8	chr10	2927289	2929289	2001	5.83	3.65	8.83	-5.19	1.53E-36	6.15E-32
9	chr3	32667457	32669457	2001	4.95	2.82	7.94	-5.12	1.72E-36	6.15E-32
10	chr12	11950970	11952970	2001	6.25	4.06	9.26	-5.21	2.87E-36	8.71E-32
11	chr2	182274990	182276990	2001	5.41	3.08	8.44	-5.36	2.97E-36	8.71E-32
12	chr13	63570473	63572473	2001	5.6	2.82	8.72	-5.9	7.36E-36	1.97E-31
13	chr1	245540753	245542753	2001	6.13	4.07	9.1	-5.03	1.46E-35	3.61E-31
14	chr3	18004789	18006789	2001	6.13	3.75	9.17	-5.42	1.68E-35	3.85E-31
15	chr10	45240790	45242790	2001	5.54	2.95	8.62	-5.67	6.63E-35	1.42E-30
16	chr21	19815181	19817181	2001	5.92	3.05	9.06	-6.01	5.21E-34	1.05E-29
17	chr3	117862996	117864996	2001	5.68	2.48	8.86	-6.37	8.60E-34	1.63E-29
18	chr16	34794174	34796174	2001	6.28	3.06	9.46	-6.4	1.08E-33	1.94E-29
19	chr8	130427019	130429019	2001	6.39	3.84	9.47	-5.63	1.95E-33	3.30E-29
20	chr14	53691496	53694404	2909	6.3	3.47	9.43	-5.96	5.56E-33	8.95E-29
21	chr11	92494436	92496436	2001	5.94	3	9.09	-6.09	9.88E-33	1.52E-28
22	chr8	85676331	85678331	2001	5.53	2.8	8.65	-5.85	1.28E-32	1.87E-28
23	chr4	142833219	142836335	3117	5.88	3.89	8.82	-4.93	2.02E-32	2.83E-28
24	chr9	93737402	93739402	2001	6.51	3.46	9.67	-6.2	7.14E-32	9.58E-28
25	chr14	26312206	26314206	2001	5.42	2.54	8.55	-6.01	1.19E-31	1.53E-27

26	chr1	31949069	31951069	2001	5.44	2.88	8.52	-5.65	1.34E-31	1.66E-27
27	chr11	16358469	16360469	2001	4.57	2.66	7.49	-4.83	2.01E-31	2.40E-27
28	chr15	86625030	86627030	2001	6.09	3.79	9.12	-5.33	3.89E-31	4.47E-27
29	chr6	7344576	7346576	2001	5	3.19	7.9	-4.71	4.03E-31	4.47E-27
30	chr6	47219564	47221564	2001	5.23	3.52	8.08	-4.56	5.93E-31	6.36E-27
31	chr8	105879304	105881304	2001	5.95	2.97	9.1	-6.13	9.89E-31	1.03E-26
32	chr16	88644554	88646554	2001	5.81	4.3	8.59	-4.29	1.13E-30	1.13E-26
33	chr20	15026936	15028936	2001	5.69	2.76	8.83	-6.07	1.65E-30	1.61E-26
34	chr3	18059502	18061502	2001	4.82	3.26	7.62	-4.36	2.74E-30	2.59E-26
35	chr3	18023441	18025441	2001	5.96	3.31	9.06	-5.76	2.98E-30	2.73E-26
36	chr6	14462591	14466197	3607	5.62	4.19	8.36	-4.16	3.05E-30	2.73E-26
37	chr11	104209396	104213028	3633	5.97	3.8	8.97	-5.17	5.39E-30	4.69E-26
38	chr3	18126796	18130734	3939	6.25	4.43	9.15	-4.72	5.71E-30	4.84E-26
39	chr15	56969960	56971960	2001	5.41	2.87	8.49	-5.62	7.79E-30	6.43E-26
40	chr7	92380738	92385495	4758	7.18	5.06	10.16	-5.11	9.69E-30	7.80E-26
41	chr2	66561018	66563018	2001	5.36	3.23	8.35	-5.12	1.38E-29	1.08E-25
42	chr1	38353178	38355178	2001	5.64	3.32	8.68	-5.35	1.40E-29	1.08E-25
43	chr13	105461803	105463803	2001	4.83	3.02	7.73	-4.71	1.44E-29	1.08E-25
44	chr3	108473746	108477456	3711	5.95	4.35	8.77	-4.42	1.60E-29	1.17E-25
45	chr15	93831197	93834357	3161	6.24	4.14	9.23	-5.09	1.86E-29	1.33E-25
46	chr7	56977661	56981018	3358	3.45	0.98	6.51	-5.53	1.93E-29	1.35E-25
47	chr2	124609933	124611933	2001	3.77	2.16	6.59	-4.43	2.25E-29	1.54E-25
48	chr8	130252550	130254550	2001	5.66	3.46	8.66	-5.2	3.82E-29	2.56E-25
49	chr6	135665372	135668977	3606	5.59	4.06	8.37	-4.31	4.58E-29	3.01E-25
50	chr13	77342780	77344780	2001	5.28	3.31	8.23	-4.92	5.01E-29	3.23E-25

7.2.3 Chapter 3 - Command Lines Used

7.2.3.a FastP Code - Python

#Create an environment
$ conda create --name FastP

# To activate this environment, use
$ conda activate FastP

# Install FastP Package
$ conda install -c bioconda fastp

# Run code to trim files
$ fastp -i untrimmedfile.fastq -o trimmedfile.fastq

Will produce a .html file and .json file -> Need to re-name the .html file to save

7.2.3.b Bowtie2 Code – Python

#Create an environment
$ conda create --name Bowtie2

# To activate this environment, use
$ conda activate Bowtie2

# Install FastP Package
$ conda install -c bioconda bowtie2

# Run code to align files (Phred score > 25)
$ bowtie2 -t -x reference genome file -U filename.fastq -S filename.sam

Will produce a .sam file in your working directory

7.2.3.c Samtools Code for .bam conversion, sorting and indexing - Python

#To install samtools without conda due to ncurses issue/incompatibility

xcode-select --install

ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"

Press RETURN to continue or any other key to abort

brew install caskroom/cask/brew-cask

brew update

brew install openssl

brew install samtools

#create a conda environment for Samtools

#Find where samtools is installed and drag contents of the /bin folder into your conda env bin folder (specifically for SAM)

#check samtools version

samtools --version

#Generate SAM to BAM file

```
samtools view -Sb filename.sam > filename.bam

#Sort BAM file
# Can also be done in IGV by igvtools option —> Tools —> igvtools —> Command:Sort —> Input BAM file —> filename.sorted.bam

samtools sort filename.bam -o filenamesorted.bam

# Generate index file (.bai) for visualising in IGV
# Can be done using samtools however did not work for me

samtools index filenamesorted.bam

# Index file generation using IGV
Tools —> igvtools —> Command:Index —> Input IGV sorted bam file —> output sorted.bai file

#Can visualise in IGV by inputting IGV sorted bam file with .bai file in the same location/file path
```

7.2.3.d DiffBind code – R programming language

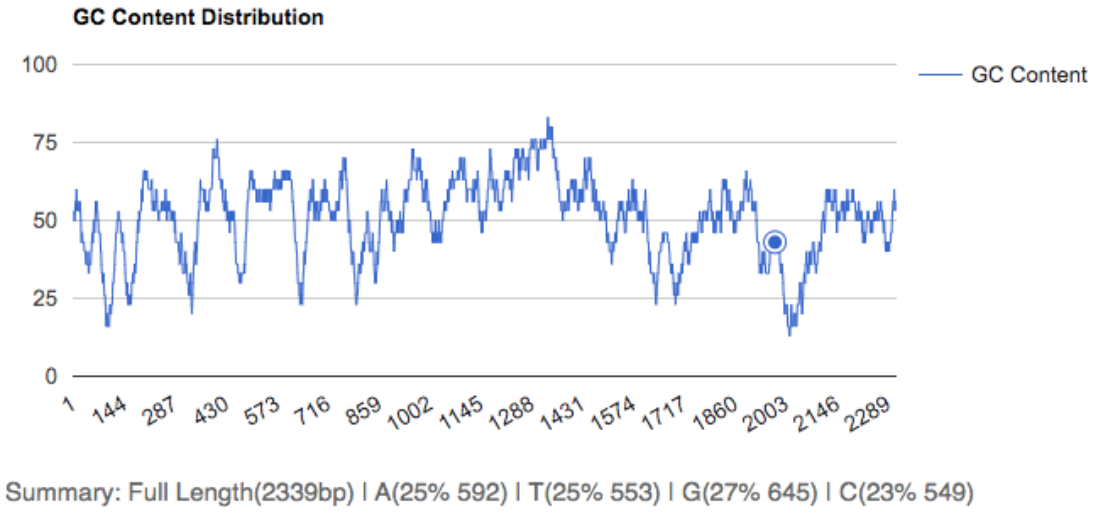
```
##All Columns should start in row A.
##Column headings are case sensitive.
##(no spaces eg. Treatment = "100 nM Dex 6hrs", should be written as "100nMDex6hrs")
##file names must match sampleSheet.csv sampleSheet.
##move files (sampleSheet.csv, reads(.bam) and peaks(.bed)) to Diffbind/extra folder.

#1) load libraries
library(DiffBind)
library(rgl)

#2) Read file path to sample sheet
samples <- read.csv(file.path(system.file("extra", package = "DiffBind"), "bed.csv"))
names(samples)
samples

#3) Create DBA from files and sample sheet
basedir <- system.file("extra", package="DiffBind")
l <- dba(sampleSheet = "bed.csv", dir=basedir)
op <- par(oma=c(5,7,1,1))
dev.off()
l

#4)Correlation heatmap, using occupancy (peak caller score) data
##Heatmap indicates correlation between the location of peaks between samples
##For large plots change the width and height when exporting
```

```
par(mar = rep(2, 4))
plot(l)

#5) Establishing a contrast between samples by tissue type
##comparison between the tissues as per sampleSheet.csv.
##minMembers refers to the number of replicates in each condition, 3 is assumed.
##We use this line to establish contrasts between replicates of the same tissue for QC of consensus peaks
##input is step 3) DBA data
DGF_peakset_Tissue <- dba.peakset(l, consensus=DBA_TISSUE)

#6) Visualise consensus peaks between replicates and calculate the % of peaks shared by at least 2 replicates
##Repeat for each sample
##QC cut-off is 75%
##If samples fail the replicate consensus QC check remove them from your sample sheet and begin at step 2) to 4) and skip to 7)
par(mfrow=c(1,1))
dba.plotVenn(DGF_peakset_Tissue, DGF_peakset_Tissue$masks$"Tissue",main="Consensus Overlaps")

#7) Create correlation heatmap, using affinity (read count) data
## trimmed mean of M (TMM) normalized (using edgeR), using ChIP read counts minus Control read counts and Full Library size
##For an in depth explanation of this normalisation method see:
https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25
##For a discussion on NGS normalisation methods see:
https://www.biorxiv.org/content/biorxiv/early/2014/06/19/006403.full.pdf
##Helpful blog: http://crazyhottommy.blogspot.com/2016/04/library-size-and-normalization-for-chip.html
##Heatmap indicates correlation between the location and signal strength of peaks (number of reads/read depth) between samples
l_count_TMM_MINUS_FULL <- dba.count(l, minOverlap=2,
score=DBA_SCORE_TMM_MINUS_FULL, summits=1000)
op <- par(oma=c(5,7,1,1))
dev.off()
l_count_TMM_MINUS_FULL
par(mar = rep(2, 4))
plot(l_count_TMM_MINUS_FULL)


#8) Differential Binding Analysis between tissues in samplesheet, using normalised read count data
l.DBAContrast_TMM_MINUS_FULL <- dba.contrast(l_count_TMM_MINUS_FULL,
categories=DBA_TISSUE, minMembers = 2)
l.DBA.analyse_TMM_MINUS_FULL <- dba.analyze(l.DBAContrast_TMM_MINUS_FULL)
op <- par(oma=c(5,7,1,1))
```

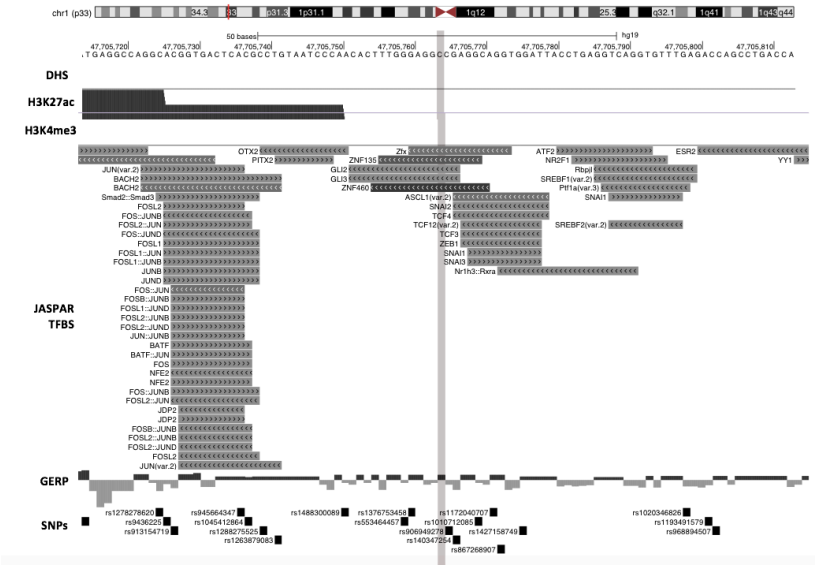
```
dev.off()
l.DBA.analyse_TMM_MINUS_FULL
par(mar = rep(2, 4))
plot(l.DBA.analyse_TMM_MINUS_FULL, contrast='X')

#9) Retrieving the differentially bound sites
##repeat for all contrasts of interest (contrast is selected for by the number in "contrast=x")
l.DB_TMM_MINUS_FULL <- dba.report(l.DBA.analyse_TMM_MINUS_FULL, contrast=1)
```

7.3 – Chapter 4: Supplementary Figures, Tables and Code

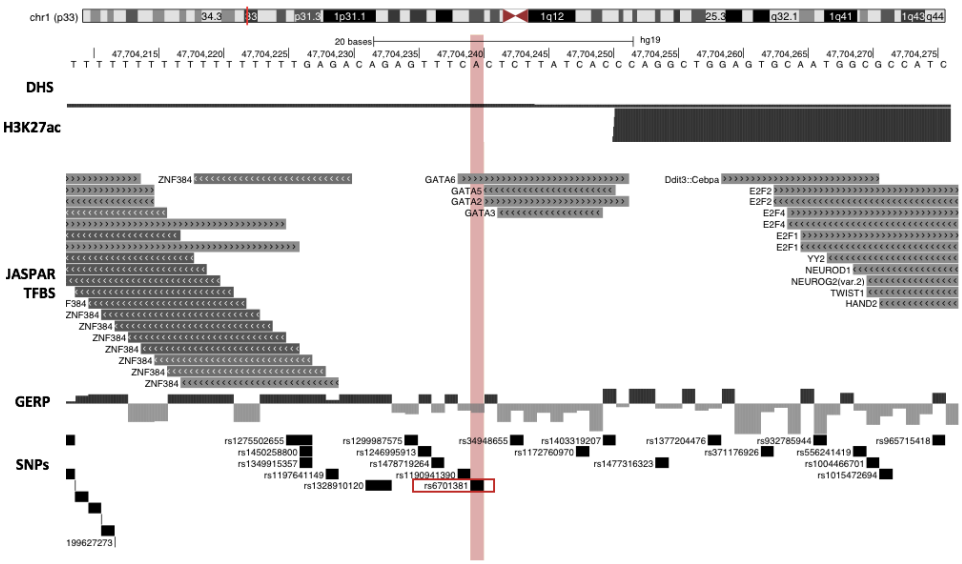
7.3.1 Supplementary Figures



| Figure 7.6. An example of GC content distribution of the sequence amplified by Seq_1 primer set within 30bp windows.
y-axis displays the percentage of GC within each 30bp window and x-axis displays bp from the start and end of the amplified sequence. Metrics of the length of the amplicon and base pair quantities and percentages overall are displayed (bottom).



| Figure 7.7. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV10 – chr1:47,705,674
(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of $< 10^{-4}$ (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.

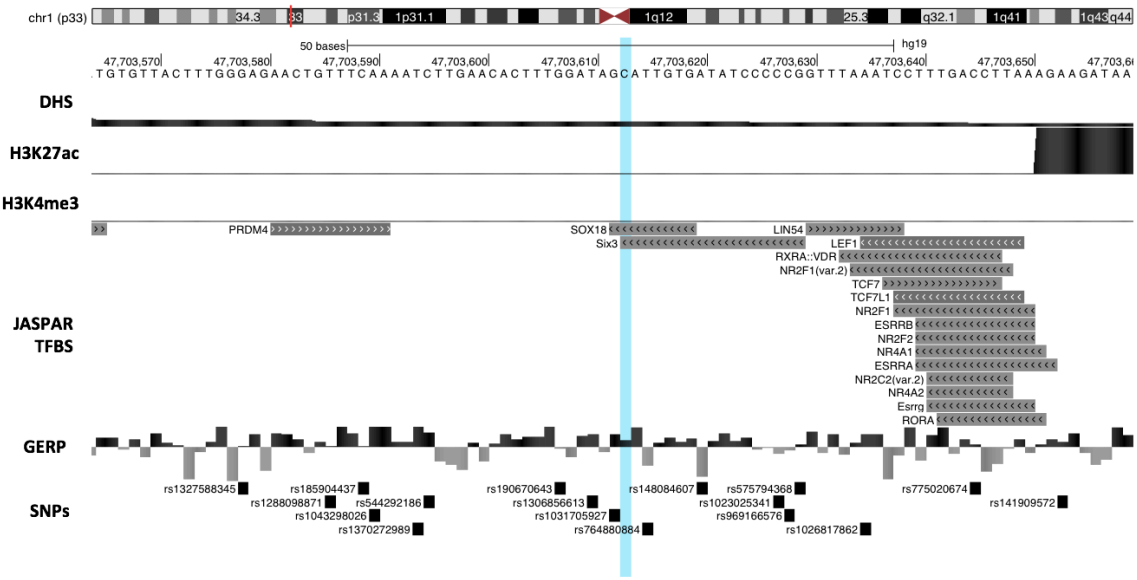


| Figure 7.8. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV8 – chr1:47,704,240

(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of $< 10^{-4}$ (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.

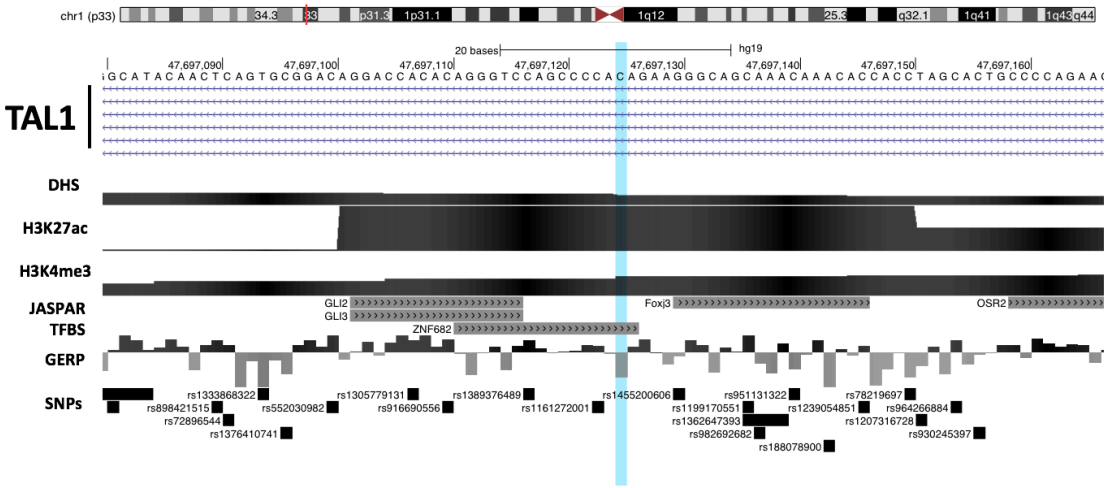
| Figure 7.9. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV9 – chr1:47,704,674

(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of $< 10^{-4}$ (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.



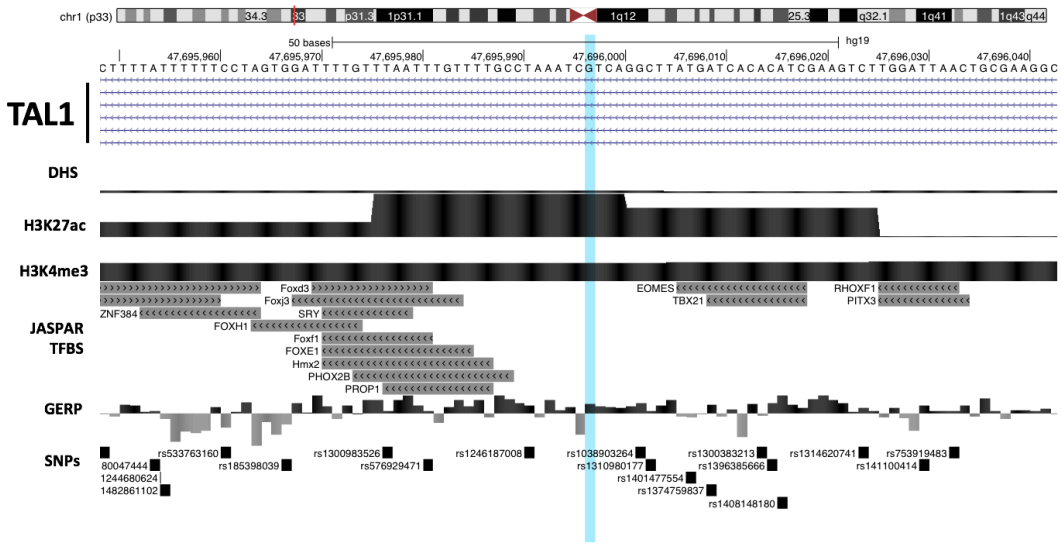
| Figure 7.10. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV7 – chr1: 47,703,613

(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of $< 10^{-4}$ (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.

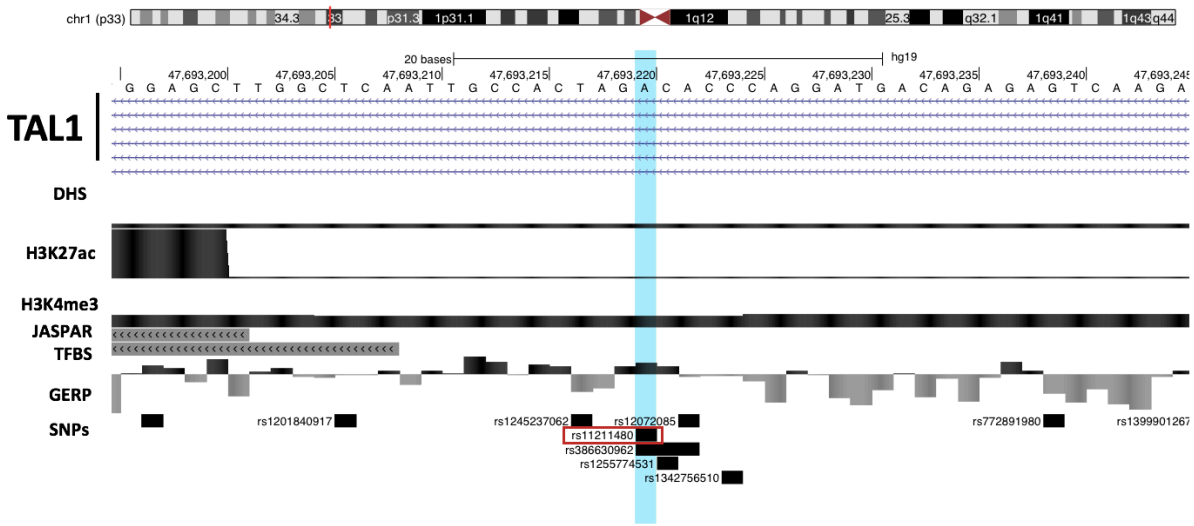


| Figure 7.11. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV6 – chr1: 47,697,125

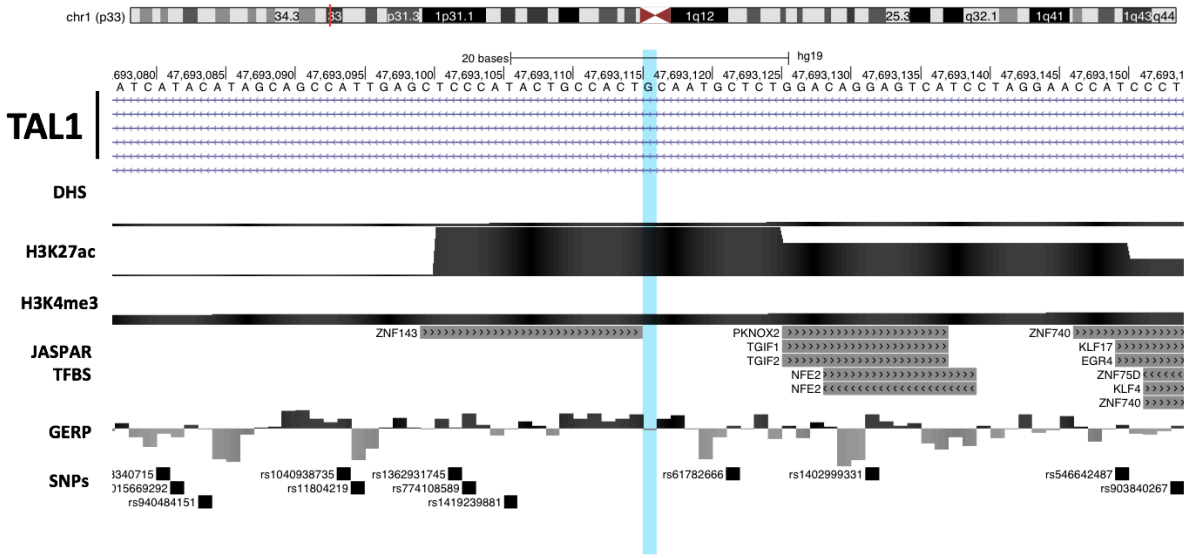
(Table 4.9) across the *TAL1* locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of $< 10^{-4}$ (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.



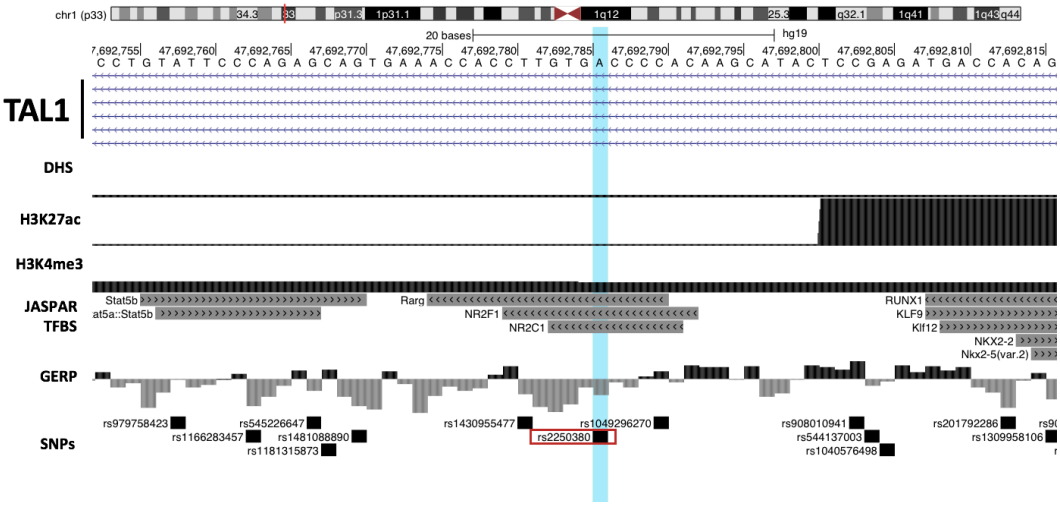
|Figure 7.12. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV5 – chr1:47,695,997
(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of $< 10^{-4}$ (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.



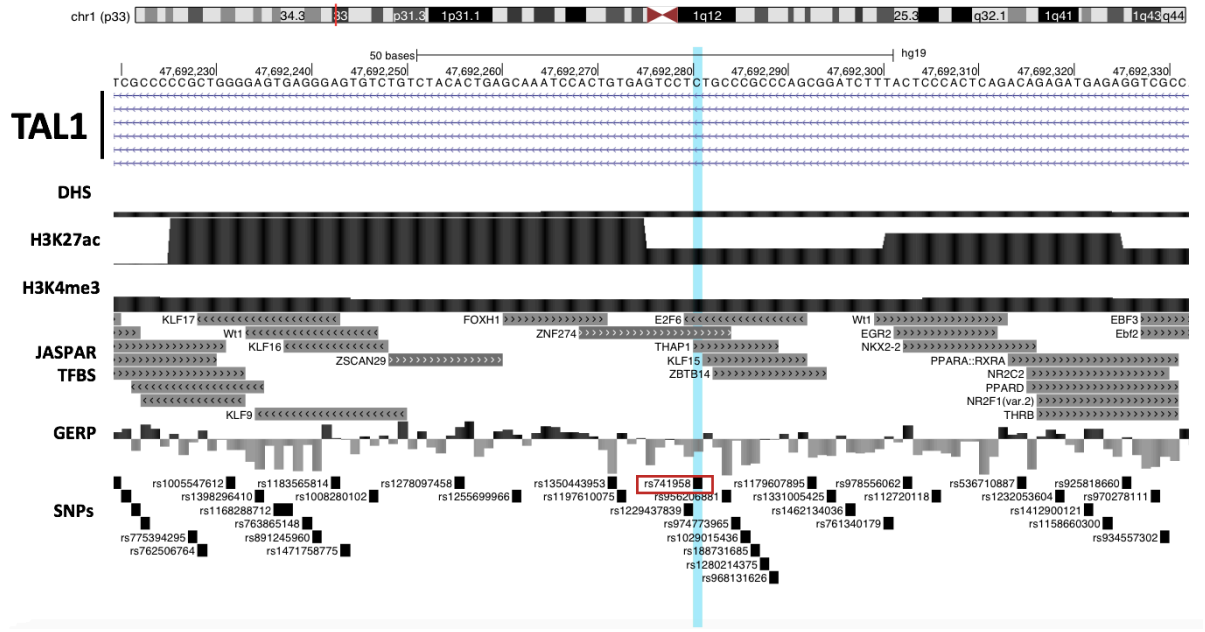
| Figure 7.13. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV4 – chr1: 47,693,220
(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of $< 10^{-4}$ (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.



| Figure 7.14. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV3 - chr1: 47,693,116
(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of < 10^{-4} (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.



| Figure 7.15. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV2 – chr1: 47,692,786
(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of < 10^{-4} (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.



| Figure 7.16. Display of predicted transcription factor binding sites (TFBS) using the JASPAR 2020 Database and mapped at Jurkat SNV1 – chr1:47,692,281
(Table 4.9) across the TAL1 locus with regulatory markers (DHS, H3K27ac and H3K4me3 – as seen in chapter 3) (Top Panel), respectively. JASPAR 2020 predicted sites display scoring of TF binding probability within a minimum p-value of < 10^{-4} (light grey) (as stated by JASPAR 2020 in the UCSC genome browser). The higher the predicted probability of TF binding is displayed the darker the shade of the TF boxes are shown.

7.3.2 Supplementary Tables

| Table 7.26. List of MSRE Primers, Primer3 outputs and sequences of MSRE primer amplicons from Chapter 4 (Methods: 4.2.1)

Primer Name	Forward Primer (5'-3')	Reverse Primer (5'-3')
MSRE_1	AGGCGGAGGATCTCATTCTT	CTAATCTCCAGGTCCCCACA
MSRE_2	CTGTCCTGAGCCTTCCTCAC	AAAAAGGGGGAAAGCAAAGA
MSRE_3	GAACATTTTCGAACCCTCCA	CTATTCGCCTTTCCCAACAC
MSRE_4	GGTTCTCCCTAAACCCCAAA	ATAAACTCGGCTGCTCATCA
MSRE_5	CGCATGTGCATTCTCTCTGT	TGCCTTGCTTCTATGGGGTA
Total Sets = 5		

7.3.2.a MSRE primer set and amplicon sequences (Methods: 4.2.1)

MSRE_1
Using 1-based sequence positions
OLIGO start len tm gc% any 3' seq
LEFT PRIMER 13 20 60.18 50.00 4.00 1.00 **AGGCGGAGGATCTCATTCTT**
RIGHT PRIMER 174 20 59.92 55.00 3.00 0.00 **CTAATCTCCAGGTCCCCACA**

SEQUENCE SIZE: 200
INCLUDED REGION SIZE: 200

PRODUCT SIZE: 162, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00

1 TACTTCATGGCCAGGCGGAGGATCTCATTCTTGCTGAGCTTCTTGTCCGGGGGATGTGTG
>>>>>>>>>>>>>>>>>>>

61 GGGATCAGCTTGCGGAGCTCGGCAAAGGCCCCGTTCACATTCTGCTGCCGCCATCGCTCC

121 CGGCTGTTGGTGAAGATACGCCGCACAACTTTGGTGTGGGGACCTGGAGATTAGGAGGAC
<<<<<<<<<<<<<<<<<<<

181 AAGAGTTAGGAGAATGGGCT

>chr1:47685696+47685857 162bp AGGCGGAGGATCTCATTCTT CTAATCTCCAGGTCCCCACA
AGGCGGAGGATCTCATTCTTgctgagcttcttgtccgggggatgtgtggg
gatcagcttgcggagctcggcaaaggccccgttcacattctgctgccgcc
atcgctccggctgttggtgaagatacgccgcacaactttggTGTGGGGA
CCTGGAGATTAG

MSRE_2

OLIGO	start	len	tm	gc%	any	3'	seq
LEFT PRIMER	40	20	59.99	60.00	4.00	1.00	**CTGTCCTGAGCCTTCCTCAC**
RIGHT PRIMER	237	20	60.05	40.00	2.00	0.00	**AAAAAGGGGGAAAGCAAAGA**

SEQUENCE SIZE: 250
INCLUDED REGION SIZE: 250

PRODUCT SIZE: 198, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00

1 TCCCCTCCCCCCACCCCCAATCTGAGAATGGGCTCTCCTCTGTCCTGAGCCTTCCTCACC
>>>>>>>>>>>>>>>>>>>

61 CCAGGCCGGAACACAAGGCCTTCCTAGACACCGTTTCCACCGGCACCATTCCCCTGAAAA

121 CTCACCTGGGGCATATTTAGAGAGACCGGCCCCTCTGAATAGGATCTCCACTCCGCCGGA

181 AAGGGGCGGAAGCCGAGGAAGAGGATGCACACCCGGGTCTTTGCTTTCCCCCTTTTTCGC
<<<<<<<<<<<<<<<<<<<

241 TGAGAGGCCT

>chr1:47694782+47694979 198bp CTGTCCTGAGCCTTCCTCAC AAAAAGGGGGAAAGCAAAGA
CTGTCCTGAGCCTTCCTCACcccaggccggaacacaaggccttcctagac

accgtttccaccggcaccattcccctgaaaactcacctggggcatattta
gagagaccggcccctctgaataggatctccactccgccggaaaggggcgg
aagccgaggaagaggatgcacaccgggTCTTTGCTTTCCCCCTTTTT

MSRE_3
OLIGO start len tm gc% any 3' seq
LEFT PRIMER 20 20 59.91 45.00 6.00 0.00 **GAACATTTTCGAACCCTCCA**
RIGHT PRIMER 189 20 59.57 50.00 2.00 0.00 **CTATTCGCCTTTCCCAACAC**
SEQUENCE SIZE: 200
INCLUDED REGION SIZE: 200

PRODUCT SIZE: 170, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00

1 CCTCCCGACAGGCTGTCTGGAACATTTTCGAACCCTCCAACTGGGATCGGTCTGGTTCAG
>>>>>>>>>>>>>>>>>>>>

61 TGTTTGTTTTCTAAGCAGGGAGGTGTCTACGCGGTTGCCTCCTCAGCCAGGTCTCCGGCT

121 GCCGCTACACCGCGAAGGGATAGTCCCGGGCCTGGATGGGCGGAGGTCCGTGTTGGGAAA
<<<<<<<<<<<

181 GGCGAATAGTCTTCAGACTC
<<<<<<<<<

>chr1:47695312+47695481 170bp GAACATTTTCGAACCCTCCA CTATTCGCCTTTCCCAACAC
GAACATTTTCGAACCCTCCAactgggatcggtctggttcagtgtttgttt
tctaagcagggaggtgtctacgcggttgcctcctcagccaggtctccggc
tgccgctacaccgcgaagggatagtcccgggcctggatgggcggaggtcc
GTGTTGGGAAAGGCGAATAG

MSRE_4
Using 1-based sequence positions
OLIGO start len tm gc% any 3' seq
LEFT PRIMER 19 20 60.16 50.00 4.00 0.00 **GGTTCTCCCTAAACCCCAAA**
RIGHT PRIMER 170 20 58.48 45.00 3.00 3.00 **ATAAACTCGGCTGCTCATCA**
SEQUENCE SIZE: 200
INCLUDED REGION SIZE: 200

PRODUCT SIZE: 152, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 0.00

1 CCCCACCACCAAAGGTTGGGTTCTCCCTAAACCCCAAAGCCAAGCCTCTTACCATTTTTC
>>>>>>>>>>>>>>>>>>>>

61 AAGTGTTTCTGCCTGCCCTCTCTGAGTACTGGACTATCTCTGAGCTTCTCCCCCGAGCTT

121 CTTCTCCGGAAAGATGCCATGCATGCACTCTGATGAGCAGCCGAGTTTATTTTTAACAAA
<<<<<<<<<<<<<<<<<<<<

181 TCTACACTGCAGTTACTGTG

>chr1:47701431+47701582 152bp GGTTCTCCCTAAACCCCAAA ATAAACTCGGCTGCTCATCA
GGTTCTCCCTAAACCCCAAAgccaagcctcttaccatttttcaagtgttt
ctgcctgccctctctgagtactggactatctctgagcttctcccccgagc
ttcttctccggaaagatgccatgcatgcactcTGATGAGCAGCCGAGTTT
AT

MSRE_5
Using 1-based sequence positions

OLIGO	start	len	tm	gc%	any	3'	seq
LEFT PRIMER	1	20	60.02	50.00	4.00	0.00	**CGCATGTGCATTCTCTCTGT**
RIGHT PRIMER	174	20	60.60	50.00	2.00	2.00	**TGCCTTGCTTCTATGGGGTA**

SEQUENCE SIZE: 174
INCLUDED REGION SIZE: 174

PRODUCT SIZE: 174, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00

1 CGCATGTGCATTCTCTCTGTCTCTAAGTCTGCTCCTCTTTCCTACCCCGGCCCTGTCTCT
>>>>>>>>>>>>>>>>>>>>

61 CCATCTCTGTCTTTAATCTGCCTCTCATCATTGCCTCCTTCCTCTTTTTTGGTCTCTGTT

121 CAGGCTGTCAATAAGAGCTCCAGCTGTGCACAGGTACCCCATAGAAGCAAGGCA
<<<<<<<<<<<<<<<<<<<<

>hg19_dna range=chr1:47705095-47705268 5'pad=0 3'pad=0 strand=+
repeatMasking=none
CGCATGTGCATTCTCTCTGTCTCTAAGTCTGCTCCTCTTTCCTACCCCGG
CCCTGTCTCTCCATCTCTGTCTTTAATCTGCCTCTCATCATTGCCTCCTT
CCTCTTTTTTGGTCTCTGTTCAGGCTGTCAATAAGAGCTCCAGCTGTGCA
CAGGTACCCCATAGAAGCAAGGCA

7.3.2.b Sequencing amplicon sequences (Methods: 4.2.5)

SEQ_1

>HG19_DNA RANGE=CHR1:47691795-47694488 5'PAD=0 3'PAD=0 STRAND=+ REPEATMASKING=NONE
CAGAGACTGAGGGCCAAAAGGACAGAGATGGAGGAAGACGACAGAGACAC
CGGAAGAAAAGGAATACAGCCAGCGACAGAAACACAGAAGGGGAAATCAG
GAGGAAGGAAATGTACAAGGAGGCAAGAAAGAGATTACTCTGTCCCCTTT
CTCAGGCCTAAAGGGAAGAGGAGGGAACAAATTCCGGATCGTGCTCTTTG

GTTTTAGCGGGGAATCTGTCAGGCAGGGGTTCGTTGCTTTAGGAACCTCC
CTTGGCCTCTCAGCAATATTCCCAGCCCACCTACAGGCACACACCCAGAA
ACCCGGTGCGGATGGAAGTAGGGAATTAGAAAGTGGAGGCTACGGGGACT
GGGATCCACTCCGCCAGAGCTGGCAGAGAATAATCCCGAGGAACCAGCTC
CCTCCACACACACACACACACACATTCGCCCCCGCTGGGGAGTGAGGGAG
TGTCTGTCTACACTGAGCAAATCCACTGTGAGTCCTCTGCCCGCCCAGCG
GATCTTTACTCCCACTCAGACAGAGATGAGAGGTCGCCAAGGGAAGAAGA
TGGTGCTGGCCCAACCTCCCTGCACAGTCCACCTCGTTCTGCTCAGCTGG
ACTCTGAAGTGAGGTCGAGTCTCAAAGAGAGGCCTGGCTGAGGTAGAATG
GAGCTGGGCATTTGAAGGGTTCCCTGAGGCCAGTGGGGTCTGGACTCCCA
TCCAAGAGCAGGACAGGATGTCATGAGGCCCTTGAACTCTTCCAGCCCCC
CTCACCGCCATCTCCTGCCTTGGTGTTCTTTGGGAAATGGGAGGAATACC
TCCTTTATGCATTTGAAAGTGCTGTGCAAACGTTCAAGTAAACATTATTA
ACAAGGAGAATGGGCCAGTGAGAAAACTGCCTGGCCGAATGTCAAGTCAC
TCACAGCCTATCTGTGAGAGAGCCAGACCAGAGCAATAATCTCACTTTGC
TCAACCCTCCTGTATTCCCAGAGCAGTGAAACCACCTTGTGACCCCACAA
GCATACTCCGAGATGACCACAGCCACTCTACTTCCCACAGGTGTGCTGGG
GCTCTTCCATCACCTGGCTTCATCCAGATAGAGCTGGGCACCAGCAAGAA
CCAAAACAGAAAAATCAGGAGACTCAAGGGCCAGAGTTTGTGAGTTGTGA
GTTGTACATATATGATCAACACCACATAAGTTGTTTAAAATAAGCTTTCT
GCTCTCTGAATTATATAATTGGGTATTTAAAAGAAGAAAACAAATACAAA
AAAACAATTTACCCCCATGCAGAGTCCATCTAAATCATACATAGCAGCCA
TTGAGCTCCCATACTGCCACTGCAATGCTCTGGACAGGAGTCATCCTAGG
AACCATCCCTCCCACCCCAGAAGGGAGAATGGATCATCTGCAGCCTCCTC
TGGAGCTTGGCTCAATTGCCACTAGACACCCAGGATGACAGAGAGTCAAG
ATTCAGACTGCTGTACACCTTCTTTGGGAATGCCAGCCCCCCAGGTCTGG
CACAGAAACCACAAAATTCCCTAAGTCCTGGTGTATACAAGCTGTAAAGG
CAGTAGGTCCCTGACACTTCATTGGAAGTGAGTTTATGATTTTGGTGGAG
ACAATAGGTCTCTATCCATAGAATGCAACCACCAAGACACCAAGGCCACT
GGAAATGCACCAGAAGTATATCAGCCAGGTCCTCTTGCTCCCAGCAGCAG
CGAAATGTGCCAGAAATATACCAACCCCTGGGAAACATACAAGCTTCAAG
GTATTGGTTCCCTACCTCACTGCTCTCCGTCCTGACACACCGGTGTGAGC
CCAGTGATCCATGCTTAAGCCCCAGCTTCTAACCGAGTGACAGTGTGCAC
TGCAGTTGCAGACTGAGCTCAAGCTCTGTTATGACACATTTTTACCCTCC
AATTAATTCTTAGGCTCCTGCTCTAGCGGTGGATTCCTGAGAGGCAATGG
GAGTACATATTCCAGATGAATTGACCCTGGCCACAAATGTCCTCTCCAGA
CACCTATAAATATCCCATCGCCAAGGTTCCTTGTAAGAAATATATGATTG
GATCCCAGGATCTGTGGAGGACCTCTCCCGGCTCCTCCCCACATTTTAGG
AAAATCCTGCCGGTTTAGGATTACCAGCCAGAGGTCCATGCCCCAAAGAA
GCCACTCACAAACACTGATGGTTGTTGGCGACTAAGCTGCATCACTTGAT
TTGGATTACAGAAAGGAACGCCCGAGAGATTTTGTTTGTTTTTATTTGGG
GAGAATGAAGGAGGAGGGAGATTTTAGACTGAATCGTTCTAGAGTATTTG
ACGACTACAGCTCCTCTCTCTTTGTACTACGGAGACCCTGCTTATAGCCC
CCAACAGGAAATCCTCATCTGCAGTTGCCAGACAGCCAGAATATTTCCAC
CTGTGCCAGAACAAGACAGAAGACCCCAGGAGATTCTTCCTGGGGACAGG
TCTCAAGCAGACCCAATCCCCCATACAGAGAATTTCTTAGATGAAAACAG
CCTCAGGCCCTTAGACCCAGCCTCTGGTCTCTCTTCTGGGCCCACACCAG
CACAACACCCTGGAGAAGCCCCAAGCCTGCACAAGTACCAAACCCACCAG
GGAACAGCACCACCATCCTCGCCTCTACCGGAGAAGAGGGGTGTTCCTGC
CTCCCCACCCTAGGCCTCAGGCCTGGGATCAGGGCCCACTCGAG

SEQ_2

>HG19_DNA RANGE=CHR1:47694309-47697743 5'PAD=0 3'PAD=0 STRAND=+ REPEATMASKING=NONE
ACCCAGCCTCTGGTCTCTCTTCTGGGCCCACACCAGCACAACACCCTGGA

GAAGCCCCAAGCCTGCACAAGTACCAAACCCACCAGGGAACAGCACCACC
ATCCTCGCCTCTACCGGAGAAGAGGGGTGTTCCTGCCTCCCCACCCTAGG
CCTCAGGCCTGGGATCAGGGCCCACTCGAGCTGGGACTGCACTAGCCAGG
AAGGCTGCCGCCTTCCTCCTGCGGGAGCTCCAGGCACAGTCTGTACCACC
CGGATACAGCCGCCCCAAAGTTACCGGCCTTGGGAAACGTCCACAGAACC
GGGACCAATGGGAAAGGCCTGGACGACTAAGTCCAACTAGCTGGTTTTAC
CCTGGGTGGAAAAGGGAAAAAGCCGCTGAATGCCACAGAGGCAGCTACAG
CAGTCAGATCCCGATTCCAAGCTCTGGCTGGTCCTCCCCTCCCCCCACCC
CCAATCTGAGAATGGGCTCTCCTCTGTCCTGAGCCTTCCTCACCCCAGGC
CGGAACACAAGGCCTTCCTAGACACCGTTTCCACCGGCACCATTCCCCTG
AAAACTCACCTGGGGCATATTTAGAGAGACCGGCCCCTCTGAATAGGATC
TCCACTCCGCCGGAAAGGGGCGGAAGCCGAGGAAGAGGATGCACACCCGG
GTCTTTGCTTTCCCCCTTTTTCGCTGAGAGGCCTGCAGTTACGCTGCGGT
GTGGTCCTGGGCGATCTGCCGCGCCCAAAGCGAGTTTCCAGGAAAGATAG
GGGTGGAGAGAAAGAGGCAGGGCAAGAGGGAGGGAGAGAGAGAAATGGGG
GTCAATGGCTGGGAATTACCTCCTGTCCCCGACCAACCAGTCCAGGGAAT
CGCAAACAGCTTTCCGCCCCCAACCTGCAGGTGTTTGGAGCCTTTCCTCT
ACCCCCTCCTCCCCCCACCCGCCTTAAAAAACCCTATACATGACCAATCA
GGAATAGCTATTTAGTCCAACAGCAACAACAACCCCTCCCGACAGGCTGT
CTGGAACATTTTCGAACCCTCCAACTGGGATCGGTCTGGTTCAGTGTTTG
TTTTCTAAGCAGGGAGGTGTCTACGCGGTTGCCTCCTCAGCCAGGTCTCC
GGCTGCCGCTACACCGCGAAGGGATAGTCCCGGGCCTGGATGGGCGGAGG
TCCGTGTTGGGAAAGGCGAATAGTCTTCAGACTCTGGTCGGTGGGACAGA
ACCTTGAAGTCCCCCCCACCTCATTCTCTCTCTGACCCCCATCCACCCAC
CTCCAGCCCAAAGCGAAACCAGAACCCAATTCCAGGGGCTGTACTTTCCT
ACAGAAATGGAGTTGGGAAGGGGGCTTGGAGAGAGATATCTGATCTGTTC
AGGTCCTTTCTCGTCTCCCTACTGATTAAGAGAAAGAAGAAAGAGAGAGG
AAGAGAGAGACAGAAAAAGAAAAGGAAATAAAGACCAACAAAGGGGGAAA
GAAAGAAACAAAGGGGAAAAGTCCCCCACCTTCCCCTCCACCACCCCTAC
GAAATGATTTTCGATTTTTCGGGGTTTAGACAGAGTAAGGGAGAATTCTA
ATGATTTGTTACAGCAATAAACAAGGTAAAAGGAAAGGAGGTAAAAACAA
ACCCGACACTACTTTCAAGGATGTAGCAGGATTGAGCTTGCTTTTATTTT
TTCCTAGTGGATTTTGTTTAATTTGTTTTGCCTAAATCGTCAGGCTTATG
ATCACACATCGAAGTCTTGGATTAACTGCGAAGGCCTCCTTCTATTTGCC
GCGGCTTTGGTGGACATATAGGAATAATTCTTCCCTGGATTGCAATACAA
TGCGGAAGATTTTCTTTCTCCTCTTCTAATCCAAAAGAGGGGAGGGGAGG
AAAGAAGAGGGGAATCCATCCATTCCCGGCGTGTTCGCGGGGGGTTAATG
TTGCGTGTTCGCTGGGGGTTAATGTTTGCCTTATGACCAAGTCTCTGTGT
CCGTGCCTCTCTCCATTTTCTCTTCCTACCTCAAACCCAGCAACTTAGAA
AACGGCTGTAGCGGAAAAAAACCCGCTGCTTTGTTCTCCCGCCAAAGGAG
CCAAAAGGGACAAACTGCTGCGCTCTGGGATGATTATTTTAATTAAAATA
GATGTTAATGATGACGATTGTGATGGTGATGGCATTAGCAATTACAATAA
TTGACAAAACAAAATTCTTTCAACCCGAGCTAGGGCCCCACTAACCCGGT
CCCTCCCAGAGCCCGGATTCGGCCTCGGTCTCGGATGCGACGGGTATCAG
ACCCAAGCGGCAACGTCGCTCACCCGAAGAGGGGGCGGGAAAATCTTCCC
AGGAGGTGATCCCGAATGTCCCCACCCGAGGCTCGAGGCCTGGTGAATGT
GCCCATTGTCCTTTCGGGGCCTCTCCCTTGGAGCCCGGCGGCCGCACCCA
CGAAACTGACCAGAAATGGATGAAGCCGGAGTGCAGCGGGAACCGAAGCC
CGCAGCTACCGGCTCGAAGGCGCCGGCCTCGGCGAGCGCTGCTCAGCTTC
AAGCCTGGAGGAGCCTCCCACTCACCAACGAACCCCTCAAACACCTAGGC
ACAGGCGGGCCTCCGGGCAGAGCAGCCGCCGACCGGGCGCTGTCCGCCCA
CCCAAGCCAACAACTGGCTCCCGAATACATCATAATTTGGAATAAAATGT
GAAATCCCGCTCCCCGCCCCCATGCCGCCGCCCCCACCAGCGCCTCGATC
TCTCGCTCGCCCTCCCCCCACTCCCGCCCCCAGCGATTTGCAAACGCACC
TCTAAAGGACACAGGCACACAGGCATACAACTCAGTGCGGACAGGACCAC
ACAGGGTCCAGCCCCACAGAAGGGCAGCAAACAAACACCACCTAGCACTG

CCCCAGAAGCCGACTTGGTCAGCCCCGCACACTGCCCAACAGGACACAAC
GAAATCAGTCAAACGCAGCGGCTCACGGACACACAATCCAGCACAGTCGG
GATCACACACGCCCGACATAACGACGACACCACCCAAACACAGTCGCAGG
GGCCACACACCCCCACACACAGAATCAGATCCCTGCTGAGAACACCAACG
GAGCACAATCGTACGCAACGAAGAAAACGCAGAAGGGCCTCGAAGGGTCC
ACATCTACACACCCCAACCGCGCAGGCACACCACACTCGGACACAGAGCC
TGTCGCCAAGAAGACCACACTTAGAAGCAGCCAACGCCGCCCACAGTTCT
CATGAACGCACTCTCACAATCCCACCGCATGCACACAACCACGAAGAAGA
AATGAAAACCAACCACAGCCTCGCGCATTTCTGTATATTGCGTAAGGAAA
AGGGGGAAGGAAGGAAGAGAGTCTCCGAGGCGGGAGGGGCGGCGGCAGCC
GGGGCGGGGGCGTCCGTGGAAAATGCCCCCCCACCGCCCCCCCCGGCCAT
CGAAAGGAACCGAGGGAAAGGGAGCGAAGACCTCT

SEQ_3

>HG19_DNA RANGE=CHR1:47697580-47701354 5'PAD=0 3'PAD=0 STRAND=+ REPEATMASKING=NONE
CGCGCATTTCTGTATATTGCGTAAGGAAAAGGGGGAAGGAAGGAAGAGAG
TCTCCGAGGCGGGAGGGGCGGCGGCAGCCGGGGCGGGGGCGTCCGTGGAA
AATGCCCCCCCACCGCCCCCCCCGGCCATCGAAAGGAACCGAGGGAAAGG
GAGCGAAGACCTCTTTTGCGGACGTGGGGAAAAAGAGGAGGAAATTGATG
AAGAATTCGGTGGGAGGCCGCGGGTTGCTTTTCCCCTAGAAAAAGGCCAG
AATGCTCACGTTTTTCCGCTACGGGGGTACGCGTGTGGCCATTTTGAGGC
CCGGTCCTGAGCGGCGGCGGCCCGGCCCCTCCCGCGGCCCCCGCTCCCCC
GCCCGCCCCGCGCCCGCCCGGCCCCTCCACCGACGCGCTGTAATCCCACT
CACGTTCCAGGCCTCGTTAGCATGGGCCGAGGCGCGCCGGGGCCGTGTGC
GCCGCAGAGATAAGGCACTGCCGCGGGTCCGCCCGCCCCCGATAAGCGCC
TCGGCCATTATGGGCCAAATGATTCATTTTAATTTGGCAATTTCACCGGA
GGGAGTGGGGACTGGTTGAGCGGCGCTAGGACCGGCTCCGGAACGCGCTG
CGGGGAGCGTTGGACGCGCTGTCTAGGACCCAGCAGATCCAGCCCCCATC
TCTAATCCCAAGGCCTCTCAGGAACCCCAGCTTCCCACAACCTCAGTCCC
TTAACCTTTCAGACACCTTTCTTCAGCGACTCTCACCCCAGCACTCTGTC
CCTACCTCTGAAGACCACCATGATACCAAAATCGTCACCTCCCAGGGCTT
CTTTCTTATTCTTCCTCCAGTGTCCCAGACACCCAGATTTCAATGGGCAG
TCTCCCTATCCCAGAAAGCTCCCCAGCCCTTCCCTACCTCTTCACCCCCT
TCTAAAATAGAGCCGGTAACTCAAATCCATGGGCTCTACGGCGGCCTGGC
AGAACCCCTGTCAATAGGGACATAAAATGCCTATTCTGCTAGGTTTCAGG
CTGAGCCCAGAGCCCAGGAACACTGCTCCCCACATGGTATTTAAAAAAAA
AAAAAAAGTTTAATCTAGCGACTGTAACACTATTCTAGGCATTCTGCAAA
TACTAATCTACTCTTCACAAACAGTCCTGAGAAGCTGGATCCTTAATTTA
CAGAAGAGAAAACTGAGGCACAGAAAGTCAAAGTAACTTTTCGAATGATA
CATAGCTAGTGCTTGGTGGAGTCTGCCTTGGAAGGAAATTTGACACGGAG
TCTTCTCTGCGGCTCAGCTGATGAAGAGGGGCAGCCAAAACCTAGCCAGG
GAAAGGGACCTGGGCTCAACTCATTGCAGAGACCAGGGTCTTTACTGTGG
GGTAGAAATCTGAGTTCAGTGCCAGCTGGATGATGCGCTCAGCCCACGGA
GGGTCTCTGGGTTCATTCTTTCATGGTGGAGTGGAAAGAACACATGTTCG
AATTTTAGTCACTGATCTCTGAACCTCAGTTTTCTCATTTGTAAAATGGG
GAGATAATGTCGACGTCACAAAGTTATGAGAACTAACTTATATGACCTTT
AAAAGGTGGTGGTGAGGATAATAATGTATGGAACGTCACTCAGTCCACAG
CCCAGGAAAACATCATATGCACTTGCATGGGAGTTCTGGTTATCATCATA
GAGAAAGAAATTAAGACACAGAGAGATGTGACTTAGCCAAAACTTACAGT
AACCCAAAGGAAAAGCCAGGACCCAAACTTTGGTGTCCAATTCCCAGTCC
ACAACTCTACTTTCCATGCCAGGCTGATTTGTCCTTCTAGTGTCTCTGCT
TTTGAGCTTTTTTTGGAGCCTGAGAACTGGGGGAAAACCCCTACTCTGGT
AACAAACTGAGTCCCAGTCAAAGACTGAGCAGAGAGCATGGCCCCAGTCT
GGGATCTGGCTTCTCCTTGGTCATCAGGTTGACCCTCTGAGAAGATACAG
TTAATTCAGTGCCGCCTGTCCCCTCTGTTCCTCCTGTGCTCTCTGAAGGA

AGTCAGGCTGTCATGGGGATACAGATAACTCTCACAGAATCCCACATCCT
AAGAAAAAGTCTAGTTCCCCCCTACCATAAGGCCTTTGCTTGGCTCCATC
ACTTATTAGAAAATGAGAGTCATATAATTCTGAGTCTGAAACACAACTCT
GCCATTCCTGGGCTGTGGGACCCTTAGAGAAGTGATTTCATGTCTCTGAG
CCACTGTAAAAACAGGCTAGCCATTCCCATATGGGAGAGCCCAGGTGGGG
ACTGCTTCTCTGAGAAAATGAGGCCTGGACTGAGATCTAAGAGTGAAATT
TGGGGGTGCGAAGAAGGGGGGTGGGGAGAAGCATCAGTATCCAGAACCTT
CCAGGTGGGCTGAGGAAACAGCTGGACCAAGGGCTAGAACAGAAGGGTGG
TAGGTTCGTGAAAAGAATAGAAAAGTGTCCAGTATGGCTGGAGCATGCTT
GTGGACAAGTTAAAAGATTTTTTGACATTTTTTTGTTTGTTTGGTTTGAA
GGTGGGGGTGGTACCTTGTATAGAATAAGATTTGCATTTCTAGAAAGTTC
TTTGGATGAAGTGTGGAGGGGGAAGGACTCTGGGAGGCCAGTTGTGGAGC
CAATGCAGTGGTTAGGGGAAGCGGATTAGAGCTTGGACCACTGTGCTGGC
AGAAGAGATGACAAGAGGCAAGTGAAGATGACATGTATTTGGGGTATAAA
ACTGGGTTTATAATAATTGGATTTGGGCCAGGCATGGTGGCTCATGCCTC
TAATCCCAGCACTTTGGGAGGCCCAGGCAGGCGGGTCACGAGGTCAAGAG
ATCGAGACCATCCTGGCCAACATGATGAAACCCTGTCTCCACTAAAAATA
CAAAAATTAGCTGGGTGTGGCAGCGCACGCCTGTAGTCTCAGCTACTCAG
GAGGCTGAGGCAGGAGAATCACCTGAACCCAGGAGGCAGAGGTTGCAGTG
AGCCGAGATCGCGACACTGCACTCCAGCCTGGCGACAGAGCGAGACTCTG
TCTCAAACAAACAAACAAAAAAAGAATTGGGTTTGGGGGGATGAGGGAGA
GGAAGGAATCCAGGAAAACCCCTGGGTTTCTGGCTGTCCTGCAGAGCTAG
TGGTGGTTATTCCTGGAGATGGGGAACCTGGAGGCTGAGGGGACATGGGG
AGGGCTTATATATTTGGTTTTGAACTCCTTAAGTTTAGATGCCTTTGAGA
CACCTGCAGACTGTTGGATTATTTTACAGGGCATTAAATGTCCCCAGCCA
GCACCTCCAAGGAAAGTGTCCTTTCCTGTGGTGAAGGAATATGGACAGGT
GGAAAGGCGTTTCTCAGCCCTGTGGCAGAGGGTGTGCCCAGGCCTCTGTC
TGTGGAGCGAGGAGGTGACGTTCACCAACCCTCCAATTCCACACATTTGT
TCTCTGCCAAGCTCCAGATTCTGACAGTCAGGCAGCCACTCCCGACCTCA
TCCACCAAAGCCCTGCCAGGGGCTGGGGTCTCATGACGTCTACATCTGCC
CCCTCCCCTGTTTCCGATGGTCAGTGGAAAAACGGAAGCTGTGCGCTAAG
CGGGGCGCTGCCTGAAATAGCATCTGGTGCCTGTCGGTCAGCGAGTTGTC
CGAGCGTCCACGAGAGAGTGACAGGCCGGGGCCAGCCAGGAAATGCCACC
CCTCCTCCACCCGCCCCAAACGGAGGCCCCAGTCGGGCAGGAAACCCGGA
GCGCCTGGGGTGGGGGTGGCGAGGAGGGAGGGGAGGTTGAGACCCAGACT
GGTAAGTCTGGAGTCTGGAACTGGG

SEQ_4

>CHR1_47699942_47702280
AGGAAACAGCTGGACCAAGGGCTAGAACAGAAGGGTGGTAGGTTCGTGAA
AAGAATAGAAAAGTGTCCAGTATGGCTGGAGCATGCTTGTGGACAAGTTA
AAAGATTTTTTGACATTTTTTTGTTTGTTTGGTTTGAAGGTGGGGGTGGT
ACCTTGTATAGAATAAGATTTGCATTTCTAGAAAGTTCTTTGGATGAAGT
GTGGAGGGGGAAGGACTCTGGGAGGCCAGTTGTGGAGCCAATGCAGTGGT
TAGGGGAAGCGGATTAGAGCTTGGACCACTGTGCTGGCAGAAGAGATGAC
AAGAGGCAAGTGAAGATGACATGTATTTGGGGTATAAAACTGGGTTTATA
ATAATTGGATTTGGGCCAGGCATGGTGGCTCATGCCTCTAATCCCAGCAC
TTTGGGAGGCCCAGGCAGGCGGGTCACGAGGTCAAGAGATCGAGACCATC
CTGGCCAACATGATGAAACCCTGTCTCCACTAAAAATACAAAAATTAGCT
GGGTGTGGCAGCGCACGCCTGTAGTCTCAGCTACTCAGGAGGCTGAGGCA
GGAGAATCACCTGAACCCAGGAGGCAGAGGTTGCAGTGAGCCGAGATCGC
GACACTGCACTCCAGCCTGGCGACAGAGCGAGACTCTGTCTCAAACAAAC
AAACAAAAAAAGAATTGGGTTTGGGGGGATGAGGGAGAGGAAGGAATCCA
GGAAAACCCCTGGGTTTCTGGCTGTCCTGCAGAGCTAGTGGTGGTTATTC

CTGGAGATGGGGAACCTGGAGGCTGAGGGGACATGGGGAGGGCTTATATA
TTTGGTTTTGAACTCCTTAAGTTTAGATGCCTTTGAGACACCTGCAGACT
GTTGGATTATTTTACAGGGCATTAAATGTCCCCAGCCAGCACCTCCAAGG
AAAGTGTCCTTTCCTGTGGTGAAGGAATATGGACAGGTGGAAAGGCGTTT
CTCAGCCCTGTGGCAGAGGGTGTGCCCAGGCCTCTGTCTGTGGAGCGAGG
AGGTGACGTTCACCAACCCTCCAATTCCACACATTTGTTCTCTGCCAAGC
TCCAGATTCTGACAGTCAGGCAGCCACTCCCGACCTCATCCACCAAAGCC
CTGCCAGGGGCTGGGGTCTCATGACGTCTACATCTGCCCCCTCCCCTGTT
TCCGATGGTCAGTGGAAAAACGGAAGCTGTGCGCTAAGCGGGGCGCTGCC
TGAAATAGCATCTGGTGCCTGTCGGTCAGCGAGTTGTCCGAGCGTCCACG
AGAGAGTGACAGGCCGGGGCCAGCCAGGAAATGCCACCCCTCCTCCACCC
GCCCCAAACGGAGGCCCCAGTCGGGCAGGAAACCCGGAGCGCCTGGGGTG
GGGGTGGCGAGGAGGGAGGGGAGGTTGAGACCCAGACTGGTAAGTCTGGA
GTCTGGAACTGGGGCAGCAAGAGCAGGGGACAAGAACAGGCTAGCTAGGC
TCTGGCCCTGAGGCCGTTGTTCCCCACCACCAAAGGTTGGGTTCTCCCTA
AACCCCAAAGCCAAGCCTCTTACCATTTTTCAAGTGTTTCTGCCTGCCCT
CTCTGAGTACTGGACTATCTCTGAGCTTCTCCCCCGAGCTTCTTCTCCGG
AAAGATGCCATGCATGCACTCTGATGAGCAGCCGAGTTTATTTTTAACAA
ATCTACACTGCAGTTACTGTGTGACAAGTACTGTCCTAGGCACTTTGTAA
GTATTAATTCAGTTAATCCTTGCAAGAACCTAATGAAGTAGGGACTACTC
CATTTTGTAGGTAGCACAGAGAGGTTGAGCAACTTGCAGAGAACACCCAG
CAGGTAAGTGGTGGAGGCAAAATTCACACCTAGGGCAGTCTGGCTTCAGG
CCACAGTGCTTCGAAAGCAGCAAAGCAAGCACAATGGTCAGGCAGACCTG
AGCTCCTATCCCTGCTCCCCTAGTTGCCAGCTGTATGATACCAGGCAAAT
TACTTAATCTCTGAGGTTCACTTTCCTTATTTGCAGAAGAGAGGTGCTCA
TTTTCCATACATAATACTGTTTATATAAGGATTATTTTTAATGAGGAAAA
TTTCAATTTGAGGAAATGTCATTATTTCCCTCTTCCACTTTCAAAGGACT
ATTCTCTCCATTTTTGCCTCTGTCTCTTGCCCTGGAAGAACCCCTACCCC
ATCAGTAAGCAGGCCTCAAGAATTCCGAGCTTGGGAGTCACGTCTGTGCT
TCCAAACCCCAACGCTGCTGTCAGAATTCTCAGGGATGGAAGGAACTTCA
CTCTCAGTGGCTCTCAAAGTGGATGGTCAGCATCACCGAGGAACCTGTTA
GAAAAGCAAATTCTCAGGCCTCACCCCAGGCCTACTACT

SEQ_5

>CHR1_47701548_47704837
GCCATGCATGCACTCTGATGAGCAGCCGAGTTTATTTTTAACAAATCTAC
ACTGCAGTTACTGTGTGACAAGTACTGTCCTAGGCACTTTGTAAGTATTA
ATTCAGTTAATCCTTGCAAGAACCTAATGAAGTAGGGACTACTCCATTTT
GTAGGTAGCACAGAGAGGTTGAGCAACTTGCAGAGAACACCCAGCAGGTA
AGTGGTGGAGGCAAAATTCACACCTAGGGCAGTCTGGCTTCAGGCCACAG
TGCTTCGAAAGCAGCAAAGCAAGCACAATGGTCAGGCAGACCTGAGCTCC
TATCCCTGCTCCCCTAGTTGCCAGCTGTATGATACCAGGCAAATTACTTA
ATCTCTGAGGTTCACTTTCCTTATTTGCAGAAGAGAGGTGCTCATTTTCC
ATACATAATACTGTTTATATAAGGATTATTTTTAATGAGGAAAATTTCAA
TTTGAGGAAATGTCATTATTTCCCTCTTCCACTTTCAAAGGACTATTCTC
TCCATTTTTGCCTCTGTCTCTTGCCCTGGAAGAACCCCTACCCCATCAGT
AAGCAGGCCTCAAGAATTCCGAGCTTGGGAGTCACGTCTGTGCTTCCAAA
CCCCAACGCTGCTGTCAGAATTCTCAGGGATGGAAGGAACTTCACTCTCA
GTGGCTCTCAAAGTGGATGGTCAGCATCACCGAGGAACCTGTTAGAAAAG
CAAATTCTCAGGCCTCACCCCAGGCCTACTACTACTACTCTGAGGGTGAG
GCCCAGCAATTTGTTTTCACAAGCCTTCCTGATGATTTCTGACACACACA
CAAAATTTGAGAACCACTTCTCTAGTTTACAACCCCCATCTGTAGCAAAG
GCCAAAGTAAAATTGTGTTAAGAAAACACCGTGCTTTTTTTTTTTTTTGA
GACAGAGTTTCACTCTGTCTCCCAGGCTGGAATACAATGGTGTGATCTCG
GCTCACTGCAACCTCTGCCTCCTGGGTTCAAGCAATTCTCCTGCCTCAGC

CTCCTGAGTAGCTGGGACTACAGGTGCATGCCGCCACACCCAGCTAATTT
TTTTTTTTTTTTTTTTTTGAGACAGAGTCTCCCTCTGTCACCCAGGCTGG
AGTGCAATGGCACAATTTTGGCTCACCACAACCTCTGCCTCATGAGTTCA
AGCAATTCTTGTGCCTCAGCCTCCCAAGAAGCTGGGACTACAGGCATGCA
TCACCACGCCCAGCTAATTACTTATATTTTTAGTAGAGACAGGGTTTTAC
CATGTTGGCCAGGCTGGTCTCGAACTCCTGACTTCAGGTGATCCGCCCGC
CTTGGCCTTCCAAAGTGCTGGGATTACAGGCATGAGCCACCACACCCGGC
CTTTTTTGTATTTTTAATAGAGATGAAGTTTCACCATGTTAGCCAGGCTG
GTCTCAAACTCCTGACCTCAGGTGATCTGCCGATCTCAGCCTCCCAAAGT
GCTGGGATTACAGGCATGAGCCACCGCCCCCGGCCTGTGCTTTTTAATGT
TTCAAAGCCATTTCCCTTTCCCTAACCCATGCGGAAGACCAAAGCATACA
GAAGACCGAAGTTGCTTTCCAGACTTACAGAAGAGATAGTGTCTAAACTT
AACTGTCCTCATGAGCAGAGAACCCCAATAGCTCTGCCATCAATGATTTT
ATTCCCAGAGCTGTTAGAAAAGGACAGAACATGTGCATAGCCACCAGTTC
TCTGGCTAGCAAATAGTCGTTGATGCTCACTGAGACAGAGTCCTGGCCCA
GGAGTACACATTAATACATGTGGACAAGACTCCTGGGCCCCTGAAGGGGA
CTCTATGTCCTGGGCACTGCCATGTGTGCCTGCACTGCTTCCTGCTTCTG
CCGCTGTTCTGTCTCTTGATCCTCCACTATAACCTGCCGGGCTGCTTGTG
TGATCCCCACTTTAAAAATGAGGAAACAGGCTCATCTCACCCAGTCACTC
ACTTGGTATACATCCAAGTCAGCTTAGCCCCGGTGTCTTCTCCATTCCTG
CTGGCTAGAATGAACTATGTGTTACTTTGGGAGAACTGTTTCAAAATCTT
GAACACTTTGGATAGCATTGTGATATCCCCCGGTTTAAATCCTTTGACCT
TAAAGAAGATAAATAGCCAAACTTATCATGAAAAGATTCTTCAACTTGTA
TTAAGGAAACAGAAATTAAAACCACAATGAGATGCCATGACACATCCACT
AAAATGATTAAAATTTAAAAGACTGACAATACCAAGTATTGCTGAGGATG
TGAAGGAGCAACTGGAGTTCTCCAGCAACATTGTTGAAGGAAATGCAAAA
ATTACTCAGCCACTTTGGAGAACTCTTTCGCAGTTCCTATGAATGTGCAG
TTTATAAACACCCATTTACTGAGACAGCTGAAAAATATATATCCACAAGA
ACACCTGTACATTAACAGTGATAGAACCCTTATTTGTAAAGCAAAAAACA
ACAATTCAACTATCCATCAACAGGTGAATGGTAAACAAATTGTGGTACAT
TCATGTAATAGAACATTACTCAGCAATAAAAAAGAAACAAATTACTGACA
AATGCCAGAATATAGACCTCTCAACAACAGGCTCAGTGAAAGAGGCCAGA
CACAAAAGACTATTTAGCATATGATTCCATTTATATGAAACTCTAGAAAA
AGCAAAAATTTTTTTTTTTTTTTTTTTTTGAGACAGAGTTTCACTCTTAT
CACCCAGGCTGGAGTGCAATGGCGCCATCTGGGCTCACGGCAACCTCCGC
CTCCTGGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGAT
TATAGGTGCCTGTCACCACTCCCAGCTAATTTTTGTATTTTTAGTAGACA
CGGAGTTTCACCATGTTGACCTCAGATGATCCGCCCGCGTCAGCCTCCCA
AAGTGCTGGGATTACAGGCATGAACCACTGTGCCCGGCCGGCAAAGCTAA
TTTCTAATGGCAAAGCTCAGATCAGGGGTGATGGAAGAGGACATATGACT
GCAAAGAGGCACAGGGAACTTTCTGCATAAAGGAAATATTCTATATCTTG
ATTGTGGTGGTGGTTATATGAGTGTATGCATTTGTCAAAACTCAACAAAC
TGTACACTAAAATGAGTACATCTTATAGTATGTAAATTATGCCTCAAAAA
ATTGACTATAAACGTTCTTTGTGGCCCCACATCAATCTTATGTTGTCCAG
GAAACCTTTTCTGATCAATACTGTCCTCAGCTGCATTTATATCTCCTCCT
CTCACCACTTGCTCTCCTGATTAGCATACCCTGGAGCCCT

SEQ_6

>CHR1_47703476_47705959
GGCTCATCTCACCCAGTCACTCACTTGGTATACATCCAAGTCAGCTTAGC
CCCGGTGTCTTCTCCATTCCTGCTGGCTAGAATGAACTATGTGTTACTTT
GGGAGAACTGTTTCAAAATCTTGAACACTTTGGATAGCATTGTGATATCC
CCCGGTTTAAATCCTTTGACCTTAAAGAAGATAAATAGCCAAACTTATCA
TGAAAAGATTCTTCAACTTGTATTAAGGAAACAGAAATTAAAACCACAAT
GAGATGCCATGACACATCCACTAAAATGATTAAAATTTAAAAGACTGACA

ATACCAAGTATTGCTGAGGATGTGAAGGAGCAACTGGAGTTCTCCAGCAA
CATTGTTGAAGGAAATGCAAAAATTACTCAGCCACTTTGGAGAACTCTTT
CGCAGTTCCTATGAATGTGCAGTTTATAAACACCCATTTACTGAGACAGC
TGAAAAATATATATCCACAAGAACACCTGTACATTAACAGTGATAGAACC
CTTATTTGTAAAGCAAAAAACAACAATTCAACTATCCATCAACAGGTGAA
TGGTAAACAAATTGTGGTACATTCATGTAATAGAACATTACTCAGCAATA
AAAAAGAAACAAATTACTGACAAATGCCAGAATATAGACCTCTCAACAAC
AGGCTCAGTGAAAGAGGCCAGACACAAAAGACTATTTAGCATATGATTCC
ATTTATATGAAACTCTAGAAAAAGCAAAAATTTTTTTTTTTTTTTTTTT
TGAGACAGAGTTTCACTCTTATCACCCAGGCTGGAGTGCAATGGCGCCAT
CTGGGCTCACGGCAACCTCCGCCTCCTGGGTTCAAGCGATTCTCCTGCCT
CAGCCTCCCAAGTAGCTGGGATTATAGGTGCCTGTCACCACTCCCAGCTA
ATTTTTGTATTTTTAGTAGACACGGAGTTTCACCATGTTGACCTCAGATG
ATCCGCCCGCGTCAGCCTCCCAAAGTGCTGGGATTACAGGCATGAACCAC
TGTGCCCGGCCGGCAAAGCTAATTTCTAATGGCAAAGCTCAGATCAGGGG
TGATGGAAGAGGACATATGACTGCAAAGAGGCACAGGGAACTTTCTGCAT
AAAGGAAATATTCTATATCTTGATTGTGGTGGTGGTTATATGAGTGTATG
CATTTGTCAAAACTCAACAAACTGTACACTAAAATGAGTACATCTTATAG
TATGTAAATTATGCCTCAAAAAATTGACTATAAACGTTCTTTGTGGCCCC
ACATCAATCTTATGTTGTCCAGGAAACCTTTTCTGATCAATACTGTCCTC
AGCTGCATTTATATCTCCTCCTCTCACCACTTGCTCTCCTGATTAGCATA
CCCTGGAGCCCTCCGTTACAGTCTAACTCATGCTGCTCAGGGCCAGGCAC
ACAGGGCACAAAAAAGGATCTGTAGACAAGGGAGGAACTGAATTAATGGT
ATTTGAAAAAGTGACAGAGACATCTGCCAGGAAGTAGGGTTACGTCTTTC
TGTGACCCTCAGTTTATCTGTAATAGGAATGGGGTGGGGCAACCACAGGA
TCTCTCTCTCCCTTTTATCTCTTATCATCTCTTTCACTCTGCTTCTCATC
ATGCCATCTCTATTTTAAGCGCATGTGCATTCTCTCTGTCTCTAAGTCTG
CTCCTCTTTCCTACCCCGGCCCTGTCTCTCCATCTCTGTCTTTAATCTGC
CTCTCATCATTGCCTCCTTCCTCTTTTTTGGTCTCTGTTCAGGCTGTCAA
TAAGAGCTCCAGCTGTGCACAGGTACCCCATAGAAGCAAGGCATGCCGAT
TCTCTTGCCACATTCCCGACAGGCTGGACCTGTTAGAAAGGCATAAGCTG
GTCTATCAGCTATATGGGACTGGGGAGAAGGGGAAATGAAAAGAAGGGTA
GAAAGGGAAGGGGAGAGATTTGGAAAGTCACCGTTCAGGAGACACACACG
CCTGGTACACTGGATGCAGGGGATTCGGCCTGCTGGACTCTGGCAGAAGC
TCTGAGTTCCCAAGAAAAGCTGTGGCCCTCTCCTCCACACGGCCCAGCCC
GTGTGCCCAGTCCAGAACCCATGCCACAGGACACAGCCTATAGCTGATTA
AGAGCACATGGTTGACTTTCACGCCACCCTGGACTAGAATGCCTCTTCAA
CAACTTAGCAGCTGATAACCCTAAGCAAGTTACATCACTTCTGTGTGCCT
CAGTTTCCTCATCTCTAAAATAGAGGTAAAAATAACACATGAGGCCAGGC
ACGGTGACTCACGCCTGTAATCCCAACACTTTGGGAGGCCGAGGCAGGTG
GATTACCTGAGGTCAGGTGTTTGAGACCAGCCTGACCAACATGGTGTAAC
CCCGTCTCTACTAAAAATAAAAAAAATTAGCCGGGCATGATGGCGTGCTC
CTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATCCCTTGAACCC
TGGAGGTGGAGGTTGCAGTGAGAAGGAGAGGCAG

| Table 7.27. Kruskal-Wallis results for DNA methylation of T-ALL, B-ALL and myeloid leukaemia cell lines for CpG Islands 1 and 2 within the *TAL1* locus.

CpG Island[1]	p-value (p<0.05)	Significant?
1	0.29	No
2	0.4	No

[1]CpG Islands tested correlate to those tested from Chapter 4 (Methods: 4.2.1) using CCLE data within the 'Methylation Plotter' program

| Table 7.28. PCR cycling conditions as conducted for Q5 high-fidelity polymerase (per manufacturer's instructions).

Cycles	Step	Temperature (°C)	Time
1	Initial Denaturation	98	30 seconds
25-35	Denaturation	98	15 seconds
	Annealing	50-72*	20 seconds
	Elongation	72	30 seconds/kb
	Extension	72	2 minutes
∞	Final Hold	10	-

* Annealing temperature is dependent on predicted annealing temperature of primer pairs (*).

| Table 7.29. Table of Proxy SNPs of tag SNPs that were found to co-localise Jurkat SNVs with linkage disequilibrium coefficient (r^2>0.8), coordinates, alleles, distance from tag SNP, r^2 coefficient, correlated alleles and RegulomeDB values.

Tag SNP	RS_Number	Coord	Alleles	Distance	R2	Correlated_Alleles	RegulomeDB
rs741958	rs741958	chr1:47692281	(C/T)	0	1	C=C,T=T	4
	rs2273834	chr1:47691090	(T/C)	-1191	0.9944	C=C,T=T	2b
	rs2250380	chr1:47692786	(A/C)	505	0.9537	C=A,T=C	5
	rs2798349	chr1:47698703	(G/A)	6422	0.9006	C=G,T=A	1b
	rs2758742	chr1:47681779	(G/T)	-10502	0.8908	C=G,T=T	4
	rs2821086	chr1:47679079	(C/G)	-13202	0.8278	C=C,T=G	4
rs2250380	rs2250380	chr1:47692786	(A/C)	0	1	A=A,C=C	5
	rs741958	chr1:47692281	(C/T)	-505	0.9537	A=C,C=T	4
	rs2273834	chr1:47691090	(T/C)	-1696	0.9483	A=C,C=T	2b
	rs2798349	chr1:47698703	(G/A)	5917	0.9444	A=G,C=A	1b
	rs2758742	chr1:47681779	(G/T)	-11007	0.9308	A=G,C=T	4
	rs2821086	chr1:47679079	(C/G)	-13707	0.865	A=C,C=G	4
	rs2249665	chr1:47687084	(A/G)	-5702	0.8383	A=A,C=G	1f
rs6701381	rs6701381	chr1:47704240	(A/G)	0	1	A=A,G=G	4
	rs35251419	chr1:47704128	(CT/-)	-112	1	A=CT,G=-	.
	rs1015890	chr1:47702229	(G/A)	-2011	1	A=G,G=A	5
	rs6700838	chr1:47700027	(C/T)	-4213	0.9982	A=C,G=T	5
	rs12057184	chr1:47706240	(C/T)	2000	0.9955	A=C,G=T	7
	rs12141363	chr1:47696581	(C/A)	-7659	0.9937	A=C,G=A	4
	rs7534271	chr1:47693981	(C/G)	-10259	0.9858	A=C,G=G	1f
	rs7525145	chr1:47693814	(G/T)	-10426	0.984	A=G,G=T	4
	rs61782665	chr1:47692035	(A/G)	-12205	0.9823	A=A,G=G	3a
	rs12407157	chr1:47707027	(G/T)	2787	0.9654	A=G,G=T	5
	rs911910	chr1:47706970	(G/A)	2730	0.9645	A=G,G=A	5

	rs2984618	chr1:47690438	(G/T)	-13802	0.9339	A=T,G=G	2b
	rs34087210	chr1:47689842	(C/G)	-14398	0.9235	A=C,G=G	4
	rs11211481	chr1:47694167	(A/G)	-10073	0.8818	A=A,G=G	1f
	rs1810658	chr1:47706733	(C/T)	2493	0.8179	A=C,G=T	7
	rs12083555	chr1:47703400	(G/T)	-840	0.8174	A=G,G=T	4
	rs11211482	chr1:47704674	(A/T)	434	0.8167	A=A,G=T	5
	rs1810659	chr1:47706731	(G/A)	2491	0.813	A=G,G=A	7
	rs977747	chr1:47684677	(T/G)	-19563	0.8093	A=T,G=G	7
	rs10789504	chr1:47676531	(T/A)	-27709	0.8026	A=T,G=A	3a
	rs741959	chr1:47676233	(A/G)	-28007	0.8026	A=A,G=G	1b
rs11211480	rs11211480	chr1:47693220	(A/G)	0	1	A=A,G=G	1f
	rs145295250	chr1:47691568	(CAGA/-)	-1652	0.958	A=CAGA,G=-	.
	rs743270	chr1:47675536	(G/C)	-17684	0.9371	A=G,G=C	4
	rs4926521	chr1:47674769	(G/A)	-18451	0.9371	A=G,G=A	5
	rs4134058	chr1:47670911	(T/C)	-22309	0.9333	A=T,G=C	7
	rs6695898	chr1:47681761	(G/A)	-11459	0.9109	A=G,G=A	4
	rs10890471	chr1:47676525	(C/A)	-16695	0.9107	A=C,G=A	4
	rs6658125	chr1:47677015	(C/T)	-16205	0.9098	A=C,G=T	6
	rs6692253	chr1:47680527	(G/A)	-12693	0.9043	A=G,G=A	5
	rs12145836	chr1:47672289	(A/T)	-20931	0.9024	A=A,G=T	6
	rs4926726	chr1:47679486	(G/A)	-13734	0.8826	A=G,G=A	5
	rs11211481	chr1:47694167	(A/G)	947	0.8684	A=A,G=G	1f
	rs11211479	chr1:47680978	(T/C)	-12242	0.8396	A=T,G=C	7
	rs4926725	chr1:47679366	(A/T)	-13854	0.8277	A=A,G=T	3a
	rs4926524	chr1:47679258	(C/T)	-13962	0.8246	A=C,G=T	4
rs11211482	rs11211482	chr1:47704674	(A/T)	0	1	A=A,T=T	5
	rs1810658	chr1:47706733	(C/T)	2059	0.9949	A=C,T=T	7
	rs12083555	chr1:47703400	(G/T)	-1274	0.9941	A=G,T=T	4
	rs1810659	chr1:47706731	(G/A)	2057	0.9932	A=G,T=A	7
	rs10890472	chr1:47708112	(G/T)	3438	0.8827	A=G,T=T	6
	rs6701381	chr1:47704240	(A/G)	-434	0.8167	A=A,T=G	4
	rs35251419	chr1:47704128	(CT/-)	-546	0.8167	A=CT,T=-	.
	rs1015890	chr1:47702229	(G/A)	-2445	0.8167	A=G,T=A	5
	rs12057184	chr1:47706240	(C/T)	1566	0.816	A=C,T=T	7
	rs6700838	chr1:47700027	(C/T)	-4647	0.8151	A=C,T=T	5
	rs12141363	chr1:47696581	(C/A)	-8093	0.813	A=C,T=A	4
	rs61782665	chr1:47692035	(A/G)	-12639	0.8064	A=A,T=G	3a
	rs7534271	chr1:47693981	(C/G)	-10693	0.805	A=C,T=G	1f
	rs7525145	chr1:47693814	(G/T)	-10860	0.8034	A=G,T=T	4

| Table 7.30. SNV3 Predicted JASPAR 2020 Transcription Factor Binding, Score and bp from start of the tested sequence for both the Jurkat alternate allele and the reference allele.

SNV3 - Alternate Allele Binding Motifs				SNV3 - Reference Allele Binding Motifs			
Name	Score	Start	End	**Name**	Score	Start	End
SOX10	8.62523	13	18	**FOXC1**	5.81625	3	10
SOX18	8.57683	12	19	**SOX18**	6.55667	12	19
FOXC1	5.81625	3	10	**NFIX**	5.47642	4	12
THAP1	7.40652	5	13	**MEIS1**	4.2416	5	11
SOX13	8.96537	11	21	**MEIS3**	5.83924	4	11
SOX15	8.0528	10	19	**NFIX**	4.79315	4	12
NFIX	5.47642	4	12	**NR2C2(var.2)**	2.14845	5	12
SOX2	8.45148	11	21	**TEAD3**	3.82584	12	19
MEIS1	4.2416	5	11	**OSR1**	6.35483	6	15
SOX8	8.48539	10	19	**NFIC**	5.16094	6	11
MEIS3	5.83924	4	11	**TEAD4**	5.81245	11	20
NFIX	4.79315	4	12	**MYB**	5.28654	6	15
NR2C2(var.2)	2.14845	5	12	**SOX18**	3.82088	7	14
NFIC	5.16094	6	11	**NKX2-8**	3.33462	8	16
NKX2-8	4.28514	8	16	**MEIS1**	2.30244	5	11
HOXD8	4.11769	7	14	**HOXA5**	4.92049	11	18
HOXB8	3.54515	7	14	**HIC2**	4.07776	5	13
FOXC1	4.50074	12	19	**THAP1**	4.86476	5	13
GSX2	2.94567	7	14	**TEAD4**	6.63411	10	21
SOX4	4.97058	11	20	**SOX18**	3.16452	13	20
MEIS1	2.30244	5	11	**NKX2-3**	3.10715	7	16
HOXA5	2.71529	7	14	**SOX18**	2.96758	8	15
HOXA7	3.24453	7	14	**SP1**	6.29953	2	11
GSX2	1.91688	13	20	**NR2C2(var.2)**	-0.513273	11	18
SOX9	6.45561	11	19	**MAFA**	8.37293	2	16
TEAD4	4.29671	11	20	**GSX2**	1.45032	13	20
HIC2	3.68843	5	13	**NKX2-2**	6.67631	5	18
HOXA4	1.83878	7	14				
MZF1(var.2)	5.96082	5	14				
NKX2-3	3.10715	7	16				
NFIA	0.958111	4	13				
SP1	6.29953	2	11				
SOX18	2.84569	9	16				
HOXB2	1.37048	7	14				
HOXB6	1.88026	7	14				
SOX10	4.9826	10	20				

| Table 7.31. SNV5 Predicted JASPAR 2020 Transcription Factor Binding, Score and bp from start of the tested sequence for both the Jurkat alternate allele and the reference allele.

SNV5 - Alternate Allele Binding Motifs				SNV5 - Reference Allele Binding Motifs			
Name	Score	Start	End	**Name**	Score	Start	End
RHOXF1	8.40591	16	23	**RHOXF1**	8.40591	16	23
HOXA4	7.1463	10	17	**MEIS1**	6.10309	11	17
BARX2	9.94126	6	17	**OTX1**	7.57923	16	23
GATA2	5.10505	12	16	**OTX2**	7.75458	16	23
OTX1	7.57923	16	23	**PITX1**	7.59749	16	23
OTX2	7.75458	16	23	**BARX2**	8.44158	6	17
PITX1	7.59749	16	23	**PITX3**	7.93544	16	24
HOXA1	5.66792	10	17	**HOXA4**	5.32326	10	17
HOXA1	5.59907	10	17	**GATA2**	4.45396	9	13
PITX3	7.93544	16	24	**NFIX**	5.12512	1	9
HOXB3	5.81558	9	18	**OTX1**	4.87037	7	14
HOXB2	5.97819	9	18	**ELK1**	7.15419	12	21
NFIX	5.12512	1	9	**GATA3**	4.70174	8	13
HOXA2	5.96118	9	18	**PITX2**	4.83765	16	23
GATA3	5.16156	8	13	**RORB**	7.62901	1	11
PBX2	9.69703	3	15	**BARHL2**	3.89466	6	15
LHX1	5.08218	10	17	**TFAP2A**	5.30024	11	19
GATA2	4.00011	9	13	**BARHL1**	3.59396	7	14
NRL	6.7159	10	20	**PAX5**	7.99346	10	21
PITX2	4.83765	16	23	**NFIA**	2.05303	1	10
HOXB4	3.65896	10	17	**GSC2**	5.38924	6	15
EVX1	5.25604	9	18	**OTX2**	3.50336	7	14
MAFK	6.38138	6	20	**BARHL2**	2.74347	5	14
RORB	7.62901	1	11	**HOXA9**	3.95523	10	19
EVX2	4.90056	9	18	**OTX2**	5.31743	14	25
DRGX	3.69116	7	14	**LHX1**	3.78116	7	14
HOXD4	3.50964	10	17	**GSC2**	4.90679	15	24
HOXC4	2.60461	10	17	**TFAP2C(var.2)**	3.61937	10	20
YY1	5.12323	10	15	**GATA5**	3.64693	6	13
MEIS1(var.2)	7.87987	3	15	**TFAP2A**	3.18783	10	20
MEIS2(var.2)	9.39649	2	16				
EN2	4.91437	9	18				
MAFF	8.63382	5	22				
JUNB	3.05497	6	16				
FOSL1::JUND	6.6337	6	15				
GATA5	4.41207	6	13				
FOSL2	3.70788	6	16				

NFIA	2.05303	1	10
LHX1	4.23992	10	17
FOXH1	3.53031	6	16
MEOX2	5.21391	9	18
MIXL1	4.45379	6	15
GATA3	4.11797	11	16
HOXB2	2.14808	10	17
ELK1	6.22605	12	21
FOXD2	4.12113	5	17
POU6F1(var.2)	4.1869	8	17
FOS::JUN	5.6358	7	13
VAX1	4.78242	10	17
EVX2	3.75618	6	15
NKX6-3	2.74893	10	18
HOXA2	3.69855	6	15
DRGX	2.32416	10	17
POU6F1(var.2)	3.94526	10	19
LHX1	3.8758	7	14
EN1	3.228	10	17
OTX2	5.31743	14	25
HAND2	2.5479	10	19
LHX9	3.99238	10	17
NKX6-2	3.98439	10	17
BARHL2	2.31102	5	14
VAX2	4.36839	10	17
GSC2	4.90679	15	24
EMX1	5.18926	9	18
HOXA7	2.88321	10	17
OTX2	2.94642	7	14
HOXB5	2.35223	10	17
FOXD2	2.58195	10	16
ZBTB18	4.32409	9	21
HOXB2	3.67298	6	15
EVX1	3.67899	6	15
TFAP4(var.2)	3.56373	10	19
JUND	-0.62119	5	15
HOXA4	1.55011	10	17
LHX6	3.16977	9	18
GSX2	1.38065	10	17
FOSL1::JUNB	6.78864	4	16
MEOX1	3.68959	9	18

PRRX1	3.90165	7	14
HOXA2	0.181628	10	17

7.3.3 Chapter 4 - Command Lines Used

7.3.3.a qcat code - Python

#Create an environment
$ conda create --name QCAT

# To activate this environment, use
$ conda activate QCAT

To install qcat
conda install -c bioconda qcat

Combine all fastq into a single file (using the *.fastq option does not seem to work with --dual)
$cat *.fastq > single_file.fastq

Detect dual barcoded samples and trim barcodes and adaptors
$qcat -f single_file.fastq -b workdirectory --trim --detect-middle

7.4.3.b Minimap2 – Python

#create an environment
$ conda create --name Minimap2

# To activate this environment
$ conda activate Minimap2

Align .fastq file to a reference genome (put into directory e.g. hg19.fa.gz)

$ minimap2 -ax map-ont hg19.fa.gz filename.fastq > filename.sam

7.4.3.c Bcftools – Python

#create an environment
$ conda create --name BCF

# To activate this environment
$ conda activateBCF

#BAM index file needs to be in the same folder as the input.bam (the index is required to find the regions specified in your .bed file)

```
$ bcftools mpileup -f hg19.fa -d 8000 -R Your_Amplicons_Regions.bed
your_alignments.sorted.bam -a
FORMAT/AD,FORMAT/ADF,FORMAT/ADR,FORMAT/DP,FORMAT/SP,INFO/AD,INFO/ADF,INF
O/ADR -Ob -o output_file.bcf

#bcftools index
bcftools index your_bcf_from_mpileup.bcf

#bcftools call --ploidy GRCh37 -Ov -R AmpliconRegions.bed -m Input.bcf -o Output.vcf
$bcftools call --ploidy GRCh37 -Ov -R your_Locus.bed.txt -m your_bcf_from_mpileup.bcf -o
your_filename.vcf
```